Visuelles Erkennen von Objekten mit Ausprägungsvarianzen

von

Jens Teichert

Dissertation

zur Erlangung des Grades eines Doktors der Ingenieurwissenschaften
(Dr.-Ing.)

Vorgelegt im Fachbereich 3 (Mathematik und Informatik)
der Universität Bremen
im September 2011

Datum des Promotionskolloquiums: 16.12.2011

Gutachter:

Prof. Dr. Rainer Malaka

Prof. Dr. Kerstin Schill

Zusammenfassung

Maschinelle visuelle Objekterkennung hat ein großes Anwendungspotential. Neben etablierten Anwendungen in der Automatisierungtechnik wären viele andere Anwendungsbereiche denkbar, wenn die Verfahren mehr von den Fähigkeiten biologischer Objekterkennung besäßen. Eine besondere Herausforderung liegt darin, dass Objekte in der realen Welt in den unterschiedlichsten Ausprägungen auftreten. Ein universelles Objekterkennungssystem muss damit umgehen können. Bisher ist nicht klar, welche Funktionsprinzipien dazu notwendig sind. Erkenntnisse über die biologische Funktionsweise sind immer noch unvollständig und erlauben keinen direkten Nachbau. Es bleibt also nur, Funktionsprinzipien anzunehmen und in Objekterkennungssystemen auszuprobieren.

In diesem Beitrag wird eine explizite Kodierung von Ausprägungen untersucht. Dazu wird ein Objekterkennungssystem erstellt, welches einige neue Verarbeitungseigenschaften besitzt. So erfolgt die Analyse von visuellem Kontext nicht wie üblich mithilfe starrer Filtermasken, sondern anhand von neu entwickelten Diffusionsverfahren. Für das Verfahren wird keine Vorverarbeitung benötigt. Es wird eine schichtenweise Repräsentation von Teilobjekten vorgenommen, die nicht wie üblich eine stufenweise Reduktion der Auflösung erfordert. Teildetektionen können von beliebigen vorausgehenden Teildetektionen abhängen und nicht wie üblich nur von der vorgeschalteten Verarbeitungsschicht.

Die Evaluierung des neu entwickelten Objekterkennungsverfahrens zeigt, dass die explizite Kodierung von Ausprägungen erfolgreich eingesetzt werden kann.

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation

1.1.1 Visuelles Erkennen

Viele Lebewesen nutzen visuelle Signale zur Beurteilung ihrer räumlichen Umgebung. Die visuellen Signale werden durch komplexe Gehirnstrukturen verarbeitet, um Repräsentationen der Umgebung zu erstellen. Diese werden mit Repräsentationen verglichen, die das Lebewesen aus Erfahrungen mit vorab aufgenommenen Signalen gewonnen hat. Dieser Vorgang wird als visuelle Wahrnehmung bezeichnet. Können die Repräsentationen als ähnlich bewertet werden, spricht man von visuellem Erkennen.

Dieser einfache Definitionsversuch findet in ähnlichen Formen eine gewisse Verbreitung in der Literatur. Aus neurophysiologischer Sicht erscheint dabei jedoch einiges fragwürdig. Sind Repräsentationen ein monolithisches Ergebnis der Verarbeitung oder sind Repräsentationen und Verarbeitung ineinander verwoben? Werden z.B. in unterschiedlichen Gehirnbereichen unterschiedliche Teil-Repräsentationen erstellt, die irgendwie zusammenwirken? Kann man die Verabeitung selbst schon als Repräsentation begreifen? Findet eine Verarbeitung der Signale tatsächlich „vor“ dem Vergleich statt - oder sind auch Verarbeitung und Vergleich ineinander verwoben? Findet also z.B. zu Beginn der Verarbeitung schon ein wenig Vergleich und zum Ende der Verarbeitung viel Vergleich statt? Gibt es überhaupt einen expliziten Vergleich oder ist das Ergebnis der Verarbeitung nur die Ableitung von Folgeverarbeitungen und -aktionen?

Die Fragestellungen machen deutlich, wie schwer eine korrekte Beschreibung visuellen Erkennens ohne Kenntnis der zugrunde liegenden Prozesse im Gehirn ist. Dort sind Verarbeitung, Repräsentationsbildung und Vergleich integriert realisiert und das Zusammenwirken ist bisher nur wenig

verstanden (Milner and Goodale, 2006). Solange kein Verständnis für die zugrunde liegenden Prozesse im Gehirn gegeben ist, orientieren sich zeitgenössische Definitionen zum Erkennen häufig an Denk- und Beschreibungsmustern, die durch zeitgenössische Technologien geprägt sind. Verarbeitungsschritte erfolgen in heutiger Technologie meist sequenziell; es gibt präzise definierte Schnittstellen und Repräsentationen, die separat analysierbar sind.

Auf psychologischer Ebene wird Erkennen oft als ein kognitiver Prozess beschrieben, der einer Wahrnehmung Konzepte zuordnet. Ein Konzept kann für eine mentale Repräsentation von Objekten oder Vorgängen stehen, die gemeinsame Eigenschaften besitzen. Die Objekte oder Vorgänge lassen sich anhand der Eigenschaften gruppieren und bilden dadurch abstrahierende Klassen. Ein Konzept steht für die Repräsentation einer Objekt- oder Vorgangsklasse.

1.1.2 Warum Objekte mit Ausprägungsvarianzen?

Visuelles Erkennen von Objekten ist schwer und benötigt umfangreiche Verarbeitungsstrukturen. Das lässt der Umstand erahnen, dass beim Menschen und vielen Tieren ein Großteil der Großhirnrinde für visuelle Informationsverarbeitung genutzt wird (Bergman and Adel, 2005). Der Grund dafür liegt sicherlich darin, dass visuelle Signale recht komplexen Varianzen unterliegen.

Im einfachsten Fall ergeben sich Positionsvarianzen und Skalierungsvarianzen dadurch, dass Objekte an unterschiedlichen Stellen im Raum angeordnet sind. Die Ausrichtung im Raum ergibt Rotationsvarianzen und perspektivische Varianzen. Je nach Beleuchtung und Oberflächeneigenschaften ergeben sich Helligkeits-, Kontrast- und Farbvarianzen. Die bisher genannten Varianzen sollen unter dem Begriff klassische Varianzen zusammengefasst werden.

Schwieriger zu verarbeiten sind topologische Varianzen, die sich aus unterschiedlichen Anordnungen von Teilobjekten ergeben. Eine Pferdeansicht verändert z.B. ihre visuellen Eigenschaften, wenn das Pferd den Kopf senkt oder die Beine in unterschiedlichen Positionen beobachtet werden. Bei unbelebten Gegenständen kann sogar die Anzahl der Teilobjekte variieren. Häuser zeigen z.B. eine variierende Anzahl von Fenstern. Fahrzeuge haben in der Seitenansicht zwei oder mehr Räder. Ebenso variiert die Ausprägung von Teilobjekten. Eine Radkappe kann z.B. unterschiedliche Herstellerlogos zeigen.

Objekte mit Ausprägungsvarianzen sind zum Beispiel Alltagsgegenstände. Es liegt eine besondere Herausforderung im Erkennen von Alltagsge-

genständen. Sie zeigen meist völlig unterschiedliche visuelle Strukturen. Was die Ausprägungen solcher Gegenstände eint, ist oft allein das vom Mensch zugeordnete Konzept, das sich meist eher nach der Funktion eines Gegenstands und nicht so sehr an dessen Aussehen orientiert. So findet man sehr viele unterschiedliche Objektausprägungen zu Konzepten wie z.B. Uhr, Schlüssel oder Schreibwerkzeug. Ausprägungsunterschiede dieser Art sollen im Folgenden „abstrakte Varianzen“ genannt werden. Auch Abbildungen von Menschen zeigen abstrakte Varianzen. So zeigen z.B. Gesten, die eine gleiche Bedeutung tragen, oft in der Anordnung der dafür eingesetzten Körperteile große Unterschiede.

Mit abstrakten Varianzen umzugehen, ist für ein Erkennungsverfahren ebenso wesentlich, wie der Umgang mit den klassischen Varianzen des Bilderkennens. Abbildung 1.1 zeigt verschiedene Varianzen für unterschiedliche Konzepte.

Konzepte, die abstrakte Abbildungsvarianzen zeigen, sind oft aus Teilkonzepten zusammengesetzt, die markant für das Konzept sind. Für das abstrakte Konzept „Stuhl“ wird man z.B. meist eine Fläche finden, auf der man sitzen kann. Ein Telefon wird meist eine Einrichtung zum Wählen von Nummern zeigen. Teilkonzepte sind oft in charakteristischen Anordnungen gegeben. So findet man z.B. für Fahrzeuge die Räder meist unten, horizontal ausgerichtet. Teilkonzepte können rekrusiv abstrakte Abbildungen zeigen, wie z.B. Anzeigen einer Uhr, die digital oder analog ausgeführt sind. Letztere können wiederum arabische oder römische Zahlen zeigen.

Der Übergang von topologischen zu abstrakten Varianzen ist stufenlos. Wären z.B. Abbildungen eines Menschen in unterschiedlichen Posen noch eine topologische Varianz, so unterlägen Bilder eines Menschen, der unterschiedliche Lasten trägt, bereits einer abstrakten Varianz. Die Abbildungsvarianzen der unterschiedlichen Posen und der Last können gemeinsam auftreten und dabei in beliebigem Ausmaß gegeben sein. Das Ausmaß von abstrakten Abbildungsvarianzen ist genauso variabel, wie das anderer Varianzen auch. Mag das Konzept des Lastenträger auf Abbildungen noch beliebig kleine abstrakte Varianzen der visuellen Signale zeigen, dann besäßen Abbildungen zum Konzept „Dschungel“ sicherlich hohe abstrakte Varianzen.

Daran gemessen, zeigen menschliche Gesten eher mittlere abstrakte Abbildungsvarianzen. Man läuft also mit dem Entwickeln von Erkennungsverfahren dafür nicht Gefahr, Verfahren zu erhalten, die im Grunde nur sehr gut im Erkennen von topologischen Varianzen sind. Auf der anderen Seite

[1] Bilder aus der Amsterdam Library of Object Images (ALOI)
`http://staff.science.uva.nl/~aloi`

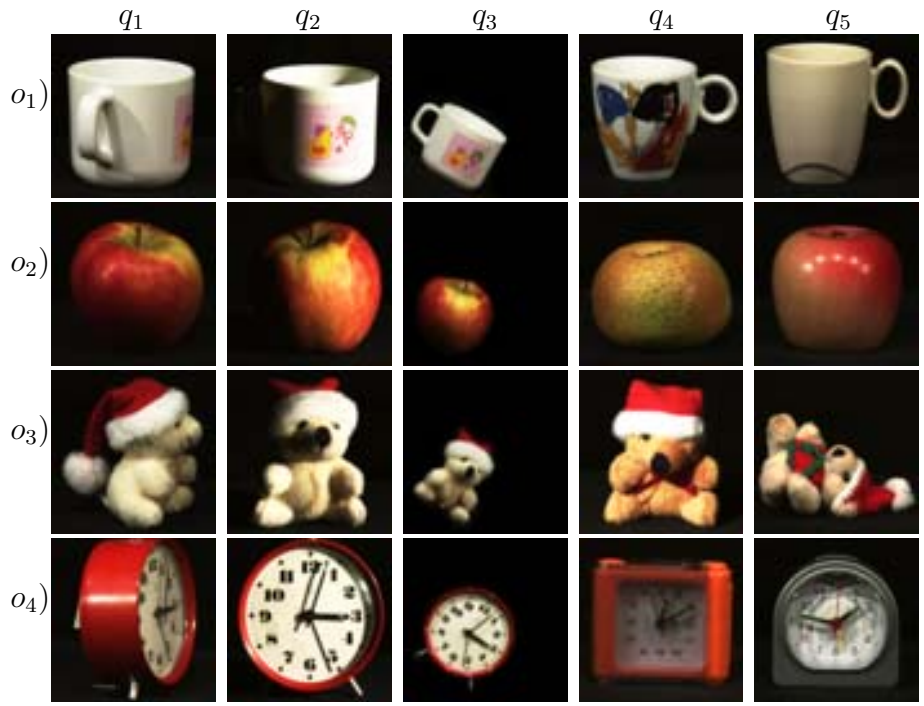


Abbildung 1.1: Für vier Konzepte o sind Muster q dargestellt, die das Konzept in verschiedenen Abbildungsausprägungen zeigen[1]. Die Muster q_1 bis q_3 zeigen Mischungen aus klassischen Varianzen. Die Muster q_4 und q_5 zeigen abstrakte Varianzen.

steigt man nicht gleich mit extrem schwierigen Beispielen ein.

Aus ähnlichen Gründen sollen auch nur zweidimensionale Bilder als Eingabe dienen. Mit zusätzlicher dreidimensionaler Sensorik wären einige der klassischen Varianzen weniger problematisch (Klette et al., 1998; Hartley and Zisserman, 2000). Man liefe aber Gefahr, Verfahren zu entwickeln, die Modelle nur gut einpassen (registrieren) können. In Realweltanwendungen unterliegen aber die dreidimensionalen Geometrien selbst abstrakten Varianzen und die aufliegenden visuellen Strukturen (Texturen) unterliegen ebenfalls abstrakten Varianzen. Das Erkennungsproblem soll erstmal im Zweidimensionalen anhand von monokularen Bildern gelöst werden, bevor Lösungen für das Erkennen im Dreidimensionalen entwickelt werden.

Das gilt auch für die Auswertung von Bewegtbildern. Aus dem optischen Fluss der Bildstrukturen können im Allgemeinen recht gut Objektumrisse gewonnen werden. Da auch diese abstrakten Varianzen unterliegen, ergibt sich kein direkter Vorteil zum Erstellen eines Erkennerverfahrens. Wenn

man ein solches aber erstmal funktionsfähig hat, liefern die Bewegtbilder wertvolle Zusatzinformationen zum Disambiguieren von dreidimensionalen Bildinhalten.

In diesem Beitrag soll es im Wesentlichen um das Erkennen von Bildstrukturen gehen und daher wird zunächst auch Farbinformation vernachlässigt. Diese trägt zwar wertvolle Information über zusammenhängende Bildregionen - die menschliche Objekterkennung funktioniert jedoch auch auf Schwarz-Weissbildern. Wenn erstmal ein funktionierendes Verfahren erstellt worden ist, kann die Auswertung von Farbinformationen später zum Verbessern der Eigenschaften integriert werden.

1.1.3 Nutzen

Das visuelle Erkennen von abstrakten Varianzen ist bisher unbefriedigend gelöst. Funktionierende Verfahren würden eine sehr wertvolle Erweiterung von Fähigkeiten künstlicher Systeme darstellen. Im industriellen Umfeld gäbe es Effizienzsteigerungen für alle Prozesse, die bisher hauptpsächlich auf die visuellen Wahrnehmungsfähigkeiten eines Menschen angewiesen sind. Die Anwendungen sind überaus zahlreich und reichen von Arbeiten am Fließband (Sortierung, Montage) bis zur visuellen Überwachung von komplexen Objekt- und Prozesseigenschaften. Roboter hätten zudem einen Vorteil davon, wenn sie ein wenig von ihrer visuellen Umgebung „verstehen“ könnten und so Ausnahmesituationen besser beurteilen und ggfs. selbst bewältigen könnten. Das selbständige Arbeiten von Maschinen in natürlichen Umgebungen könnte verbessert bzw. erst ermöglicht werden. So könnten landwirtschaftliche Maschinen autonomer pflegen und ernten. Die Beurteilung von Straßenszenen durch Fahrassistenzsysteme könnte deutlich sicherer werden. Immer mehr Automaten werden zur Rationalisierung von Kundenabfertigung im Verkauf, Autorisierung und Service eingesetzt. Die Qualität der geführten Dialoge könnte deutlich gesteigert werden, wenn einer Maschine Informationen über den visuellen Kontext des Benutzers zur Verfügung stehen würde.

Heutige Verfahren zum visuellen Erkennen sind gemessen an menschlichen Fähigkeiten immer noch sehr beschränkt in ihrer Leistungsfähigkeit. Ein Verfahren zum Erkennen von abstrakten Varianzen ist sicherlich ein wichtiger Schritt auf dem Weg zu einem universellen Erkennungssystem.

Im menschlichen Gehirn werden neben visuellen Signalen auch noch andere Modalitäten durch ähnliche Gehirnstrukturen verarbeitet. Daher ist die Hoffnung berechtigt, durch ein besseres Verständnis und die Möglichkeit zur Nachahmung visueller Fähigkeiten auch die Verarbeitung anderer Modalitäten, wie Sprachverstehen oder Bewegungssteuerung besser nachahmen zu

können. Gegenüber anderen Modalitäten hat man bei der Erforschung der visuellen Informationsverarbeitung wohl einfacher die Möglichkeiten, auch Zwischenschritte zu untersuchen. So ist die Detektion von Teilkonzepten für visuelle Signale wahrscheinlich besser zu veranschaulichen und zu bewerten, als bei akustischen Signalen. Insofern stellen Verfahren zum visuellen Verstehen auch Schlüsseltechnologien für andere Modalitäten dar. Sind erstmal gundlegende Herausforderungen der Wahrnehmung gelöst, wird das auch ein wesentlicher Schlüssel zum Verständnis weiterer kognitiver Fähigkeiten und deren Modellierung sein.

1.2 Problemstellung

1.2.1 Universelles Objekterkennen

Es soll ein Verfahren zum visuellen Erkennen angestrebt werden, das in der Lage ist, Abbildungen mit abstrakten Varianzen zu erkennen. Dabei kann es sich z.B. um Schreibwerkzeuge, Uhren oder auch menschliche Gesten handeln.

Die Erkennung soll allein aus den visuellen Strukturen der Abbildung erfolgen. Es sollen also keine weiteren Sensordaten verwendet werden, die z.B. Informationen zur Raumtiefe aus Stereokameras oder Laserscanneren gewinnen.

Es sollen keine Einschränkungen bezüglich der Objekteigenschaften des Hintergrundes oder der Perspektive gemacht werden müssen. Dabei wäre die Gefahr gegeben, nur ein weiteres (hoch-)spezialisiertes Verfahren zur Objekterkennung zu erstellen (Alexe et al., 2010).

Es soll versucht werden, einen Schritt in Richtung eines universellen Verfahrens zum Objekterkennen zu gehen.

Ein Verfahren zum universellen Objekterkennen kann Objektabbildungen Konzepte zuordnen, so wie es der menschliche Wahrnehmungsapparat auch kann. Die Abbildung der Objekte können zu beliebigen Konzepten gehören und vor beliebigen Hintergründen gegeben sein. Die Abbildungen dürfen klassischen und abstrakten Varianzen unterliegen. D.h. ein solches System muss auch bezüglich der Varianzen zwischen gelernten Abbildungen generalisieren können.

1.2.2 Biologisches Vorbild

Universelles Objekterkennen erfordert Verfahren, die abstrakte Varianzen verarbeiten können. Bisher sind solche nur in biologischen Systemen reali-

siert. Sie dienen als Vorgabe der zu erreichenden Funktionalität.

Durch Analyse der biologischen Verarbeitungsstrukturen ist es bisher nicht gelungen, die wesentlichen zugrunde liegenden Verarbeitungsprinzipien aufzudecken. Die Verarbeitungsstrukturen sind jedoch so aufgebaut, dass sie sich im Gewebe räumlich wiederholen und auch für andere Sinnesmodalitäten ähnlich sind. So erscheint die Hoffnung berechtigt, dass es für die Verarbeitung Prinzipien gibt, die allgemein anwendbar sind.

Neben Informationsverarbeitung sind im Gewebe auch viele andere Funktionen z.B. Stoffwechsel- und Immunfunktionen integriert. Bisher ist nicht klar zuzuordnen, welche Komponenten welche Rolle für welche Funktionen spielen. Unklar ist auch, wie die Integration von verschiedenen Funktionen von z.B. Bewegungserkennung, Tiefenwahrnehmung, Lernen oder visuelles Vorstellen realisiert ist. Bisher ist es nicht gelungen, einen Bogen zu spannen zwischen den bekannten physikalischen sowie biochemischen Fakten und den Verarbeitungsprinzipien solcher Wahrnehmungsfunktionen.

Es kann im Rahmen dieses Beitrags nicht versucht werden, Wahrnehmungsfunktionen durch Simulation der bekannten Physiologie nachzubilden. Es bleibt groß angelegten Forschungsprojekten[2] vorbehalten, diesen Weg zu gehen und aus solchen Simulationen, Hyopthesen zu Verarbeitungsprinzipien abzuleiten.

Fraglich bleibt, ob Phänomene wie Wahrnehmung oder Bewusstsein überhaupt durch Simulationen mit zeitgenössischen Computern zu erreichen sind. Im Rahmen dieses Beitrags soll davon ausgegangen werden, dass wenigstens Wahrnehmungsfunktionen nachzubilden sind.

1.2.3 Herangehensweise

Für diesen Beitrag steht die Umsetzung von hypothetischen biologischen Verarbeitungsprinzipen im Vordergrund. Dabei soll nicht versucht werden, die biologischen Vorgänge im Detail nachzubilden, sondern nur deren abstrakte Funktionsweise.

Wie in wissenschaftlichen Ansätzen üblich, sollen in diesem Beitrag Hypothesen über Verarbeitungsprinzipien postuliert und anschließend evaluiert werden, welche Relevanz diese Hypothesen für die gewünschte Funktionalität hat. Die Wahrnehmungsforschung profitiert von einem Wechselspiel aus physiologischer Forschung und den Ergebnissen solcher konstruktiver Simulationen von Verarbeitungsprinzipien. Diese Iteration lenkt beide Forschungsdisziplinen auf relevante Fragestellungen, die im Ergebnis (hoffentlich) effiziente Verarbeitungsprinzipien aufdecken werden.

[2]Das Blue Brain Project `http://bluebrain.epfl.ch`

1.2.4 Wissenschaftliche Fragestellung

In diesem Beitrag soll ein Weg gefunden werden, wichtige Funktionsprinzipien biologischen Bilderkennens in ein Verfahren zu integrieren. Welche Funktionsprinzipien erscheinen dazu untersuchenswert? Wie können diese realisiert werden, so dass ein Zusammenwirken möglich wird?

1.3 Gliederung der Arbeit

Kapitel 2 beschreibt wie visuelles Erkennen modelliert werden kann. Dazu werden zunächst Herausfoderungen formuliert, die im Folgenden als Gliederung verwendet werden.

Danach werden relevante Erkenntnisse über das biologische Vorbild dargestellt. Dabei werden auch Arbeiten berücksichtigt, Hypothesen über die mögliche biologische Funktionsweisen an künstlichen Systemen verifizieren. In Kapitel 3 werden dann Verfahren zum Bilderkennen beschrieben, deren Eigenschaften im Sinne der wissenschaftlichen Fragestellung interessant genug erscheinen.

Um dies systematisch durchzuführen, wird wieder die in Kapitel 2 eingeführte Gliederung verwendet. Diese wird auch verwendet, um in Kapitel 4 in die Eigenschaften des neu erstellten Provadero-Verfahrens einzuführen. Dessen detaillierte Realisierung wird in Kapitel 5 beschrieben. In Kapitel 6 wird über Evaluierungen des Provadero-Verfahrens berichtet. Zunächst wird ein Diffusionsverfahren des Verfahrens einzeln betrachtet und dann das Gesamtsystem. Dabei werden erst Tests zu klassischen und dann zu abstrakten Varianzen auf Realweltbildern durchgeführt. Kapitel 7 bringt eine abschließende Zusammenfassung und einen Ausblick.

Kapitel 2

Visuelles Erkennen modellieren

In diesem Kapitel soll im Rahmen der wissenschatlichen Fragestellung relevantes Wissen aus den Forschungsdisziplinen Anatomie, Biophysik und Wahrnehmungspsychologie über biologische Wahrnehmungssysteme zusammengetragen werden.

Zunächst werden Fragestellung benannt, die sich in der Auseinandersetzung mit der wissenschaftlichen Fragestellung als zentrale Probleme erwiesen haben. Für jedes universelle Objekterkennungssystem müssen Lösungen für diese Fragestellungen gefunden werden.

2.1 Herausforderungen

Biologische Wahrnehmungssysteme besitzen eine bisher nicht handhabbare Komplexität. Sie entsprechen in ihrem Aufbau nicht den bisher vom Menschen erstellten Technologien und sind daher auch mit etablierten Denkmustern nur schwer zu analysieren. Beides macht es schwer, zugrundeliegende Funktionsprinzipien zu ermitteln und in ausreichend funktionierende Modellsysteme umzusetzen. Aus technischer Sicht bleibt das Problem, Objekte in visuellen Szenen zu erkennen schon seit einigen Jahrzehnte ein „Heiliger Gral“ (Mel and Fiser, 2000). Die Frage, ob es überhaupt möglich ist, biologische Wahrnehmungssysteme nachzubauen wird wohl erst zu beantworten sein, wenn dies tatsächlich gelungen ist.

In der Auseinandersetzung mit der wissenschaftlichen Fragestellung lassen sich einige zentrale Fragestellungen identifizieren. Die Aufteilung der Fragen ist in diesem Sinn als willkürlich anzusehen. Sie werden im Rahmen dieser Arbeit nur zur Gliederung der Inhalte in verschiedenen Kapiteln eingesetzt.

Übergeordnet ist die Frage:

- Wie kann ein biologisches Wahrnehmungssystem auf eine künstliche technische Struktur abgebildet werden?

2.1.1 Effizienz

Biologische Wahrnehmungssysteme helfen ihren Individuen, ihr Überleben zu sichern. Aufgaben von Wahrnehmungssysteme sind z.B. Nahrung, Feinde oder Fortpflanzungspartner zu erkennen und zu unterscheiden. Für technische Systeme können die Aufgaben weniger anspruchsvoll sein. Wichtig ist jedoch meist, dass die Funktion gesichert ist:

- Wie kann sichergestellt werden, dass ein künstliches Wahrnehmungssystem eine bestimmte Aufgabe erfüllt?

Natürlich sind auch Fragestellungen nach Effizienz im klassischen Sinn wichtig:

- Wie kann ein künstliches Wahrnehmungssystem mit möglichst geringer Komplexität der abbildenden technischen Struktur realisiert werden?
- Wie kann ein Wahrnehmungsvorgang mit möglichst geringem Verarbeitungsaufwand der abbildenden technischen Struktur realisiert werden?

2.1.2 Repräsentation

Biologische Wahrnehmungssysteme erstellen eine interne Repräsentation einer visuellen Szene. Bisher ist unklar, welchen Anforderungen diese Repräsentation genügt. Sicherlich werden für eine bewusste Wahrnehmung detaillierte interne Repräsentationen und insbesondere Konzeptualisierungen von Szenenobjekten benötigt. Dabei spielt wahrscheinlich auch eine Rolle, dass das bisher erlernte Wissen über die Objekte verteilt repräsentiert ist und ein Vergleich und damit auch die Repräsentation des betrachteten Bildes verteilt vorgenommen wird. Für einfache Individuen und technische Systeme genügen möglicherweise aber auch einfache interne Repräsentationen, die gerade nur zum Ableiten von Aktionen benötigt werden. Es bleibt die Frage:

- Was muss überhaupt repräsentiert werden?

Visuelle Signale sind typischerweise zweidimensional und zeitveränderlich. Die Zweidimensionalität stellt ein großes Problem für die Verarbeitung dar.

In künstlichen Systemen, lassen sich die Bildinformationen zwar noch gut in zweidimensionalen Matrizen von skalaren Farbinformationen repräsentieren. Die Auswertung auf diesen Repräsentationen ist jedoch schwierig, wenn Abbildungen projektiven und anderen Varianzen unterliegen. Biologische Wahrnehmungssysteme repräsentieren das zweidimensionale Bild schon bei der Aufnahme als Ergebnis lokaler und regionaler Filter. Entlang des Verarbeitungspfades verlieren dieses Teilrepräsentationen an Ortsbezug, der am Anfang des Pfades noch fest durch die Position des Filters (Neurons) im umgebenden Netzwerk gegeben ist. Später sind Information über Positionen, Strukturen und Anordnungen im Zweidimensionalen nicht mehr durch das Netzwerk gegeben. Wie diese Informationen dann kodiert sind, ist noch unklar.

- Wie werden Strukturen und Anordnungen in der Bildebene kodiert?

Schließlich muss die rekonstruierte dreidimensionale Welt repräsentiert werden. Hierbei ist unklar, wie die Positionen und Anordnungen der dreidimensionalen Szene kodiert werden. Man kann die letzte Frage verallgemeinern:

- Wie werden Strukturen und Anordnungen kodiert?

2.1.3 Integration von Repräsentationen

Wenn entlang eines Verarbeitungspfades aus Teilrepräsentationen steigend abstraktere Repräsentationen erstellt werden sollen, stellt sich die Frage, wie die Teilrepräsentationen zusammenwirken, um Folgerepräsentationen zu erstellen.

- Wie werden Repräsentationen integriert?

Beim Integrieren spielt Kontext eine wesentliche Rolle. Zu Beginn des Verarbeitungspfades sind das noch Nachbarschaften in der Bildebene. Später sind es dann eher semantische Nachbarschaften, die ausgewertet werden.

- Wie wird Kontext ausgewertet?

Realweltobjekte haben meist viele Eigenschaften, die in unterschiedlichen Kombinationen auftreten können. Es wäre nicht effizient, alle möglichen Kombinationen separat zu repräsentieren. Wahrscheinlich werden die Teilrepräsentationen, die Eigenschaften kodieren irgendwie an Objektrepräsentationen gebunden. Es gibt verschiedene Hypothesen, wie das geschehen kann.

- Wie werden Eigenschaften an Repräsentationen gebunden?

Es gibt verschiedene Möglichkeiten, wie erlerntes Wissen in den Erkennungsvorgang einwirken kann. Dies kann durch Verstärkung und Abschwächung der Verbindungen entlang des Verarbeitungspfades erfolgen. Es gibt auch sehr viele rückwirkende Verbindungen und vielfältige andere gegenseitige Beeinflussung bei der Signalweiterreichung. Welche Prinzipien hierbei wirken, ist noch nicht ermittelt.

- Wie fließt bekanntes Wissen in den Erkennungsvorgang ein?

2.1.4 Lernen

Sicherlich unterliegen viele Funktionen der Verarbeitungselemente Lernvorgängen, die o.g. Erkennungsleistungen erst möglich machen. Für die Wirkungsweise von Lernvorgängen gibt es bisher zwar einige detaillierte biophysische und -chemische Erkenntnisse. Diese können aber mangels Kenntnis über die allgemeine Verarbeitung kaum auf dieser Ebene als Prinzipien ausgedrückt werden. Daher bleibt an dieser Stelle die Frage allgemein:

- Wie funktioniert Lernen in Objekterkennungssystemen?

2.1.5 Varianzen

Eine große Herausforderung bei der visuellen Wahrnehmung ist die Objekterkennung unter klassischen Varianzen, wie Veränderungen von Position, Skalierung, Rotation, Perspektive oder Beleuchtung. Diese Varianzen sind nicht unabhängig voneinander auflösbar, weil sie immer auch die Abbildung eines Objektes und damit seine zweidimensionale Struktur zum Teil nichtlinear verändern.

- Wie funktioniert Erkennen unter klassischen Varianzen?

Schwieriger noch ist die Erkennung unter abstrakten Varianzen. Dies ist nur möglich, wenn Wissen über abstrakte Gemeinsamkeiten der Objekte einfließen kann. Die folgende Frage hat als Teil der wissenschaftlichen Fragestellung in diesem Beitrag besondere Bedeutung:

- Wie funktioniert Erkennen unter abstrakten Varianzen?

2.2 Visuelles Erkennen - Stand der Forschung

Biologische Bilderkennung lässt sich nicht einfach in symbolischer Form beschreiben. Daher fällt bisher eine Umsetzung auf Computern schwer. Um das Problem in symbolischer Form zu beschreiben oder neue Maschinen nach biologischem Vorbild zu bauen, ist zunächst ein grundlegendes Verständnis biologischer Bilderkennung erforderlich.

Dies ist aber schwer zu erlangen, da die beteiligte Biophysik eine nicht handhabbare Komplexität besitzt. Es besteht schon Unsicherheit zu bestimmen, ob eine biologische Gegebenheit tatsächlich zur Mustererkennung oder generell Informationsverarbeitung beiträgt oder, ob sie ganz anderen Zwecken wie Stoffwechsel oder Immunsystem dient. Möglicherweise sind diese Funktionen auch nicht eindeutig trennbar.

Ein anderes Problem ist die extreme Vernetzung der informationsverarbeitenden Nervenzellen. Typischerweise findet man eine starke Rückkopplung auf vorgeschaltete Zellen, was eine Analyse von Ursache und Wirkung stark erschwert. Zusätzlich wird die Analyse dadurch erschwert, dass es nichtlineare Effekte gibt und Abhängigkeiten von Signalaktivierungen auf verschiedenen Zeitskalen eine Rolle spielen. Bisher konnten Erkenntnisse aus Biophysik, Anatomie und Psychologie nicht zu einer Theorie zusammengetragen werden, die den Nachbau eines biologischen Mustererkenners ermöglichen würde.

Es besteht aber berechtigte Hoffnung, dass eine solche Theorie eines Tages gefunden werden kann. Ein Hinweis darauf ist, dass sich Gewebestrukturen im Gehirn lokal und regional wiederholen. Wenn also gleiche Verarbeitungsmodule für die Verarbeitung auch unterschiedlicher Sinnemodalitäten verwendet werden, liegt wahrscheinlich ein einheitliches Verarbeitungsprinzip vor. Ein anderer Hinweis ist, dass sich visuelle Systeme in unterschiedlichen Evolutionären Zweigen finden lassen. Diese sind oft sehr unterschiedlich gebaut, ermöglichen jedoch gleiche Fähigkeiten (konvergente Evolution). Man findet z.B. unterschiedliche visuelle Systeme bei Insekten und Säugetieren. Diese Diversität der „Implementierung“ bei ähnlicher Funktionalität gibt Hoffnung auf die Existenz eines abstrakten Verarbeitungsprinzips für visuelle Wahrnehmung.

Die Biologie implementiert dies Verarbeitungsprinzip mit einer hochvernetzten dreidimensionalen Zellstruktur und zeitabhängiger Signalverarbeitung. Andere Implementierungen könnten möglich sein. Möglicherweise gelingen anschaulichere Systeme, wenn die Verarbeitung nicht mehr an eine dreidimensionale Materie gebunden ist, sondern virtuell auf einem geeigne-

ten künstlichen System durchgeführt werden kann.

Im folgenden Abschnitt werden funktionale Prinzipien biologischer Bildverarbeitung beschrieben. Diese werden aus Forschungserkenntnissen aus Anatomie, Biophysik und Wahrnehmungspsychiologie abgeleitet. Es werden nur Prinzipien beschrieben, die zur Kategorisierung bestehender Verfahren und zur Entwicklung eines neuen Verfahrens relevant sind. Darüber hinausgehende Zuordnungen von Erkenntnisse zu funktionalen Prinzipien werden in der Literatur diskutiert (z.B. Oram and Perrett, 1994; Milner and Goodale, 1995; Tanaka, 1996; Mountcastle, 1997).

2.2.1 Ziele

Ein primitives Lebewesen extrahiert typischerweise Information aus seiner visuellen Umgebung, um zwischen Bedrohungen, Nahrung oder Fortpflanzungspartnern zu unterscheiden. Diese Unterscheidung ist schwer, weil schon die Signale von visuellen Sensoren kleiner Lebewesen einen großen Merkmalsraum aufspannen. Dieser unterliegt Transformationen, z.B. aus Beleuchtungsvariation. Die Unterscheidung kann daher nicht einfach aus einer starren Unterteilung dieses Raumes gewonnen werden. Ein anderes Problem ist, dass Realweltobjekte, die zu einer unterscheidenden Klasse gehören, meist in unterschiedlichen Bereichen im Merkmalsraum liegen. Das liegt typischerweise daran, dass Realweltobjekte starken Ausprägungsvarianzen unterliegen und so die aus dem Sensorsystem gewonnen Signale kaum mit denen der Musterklassen korrelieren.

Höher entwickelte Lebewesen müssen visuelle Information detaillierter klassifizieren können, z.B. um Werkzeuge erkennen und anwenden oder Bewegungen abschätzen zu können. Ihre visuellen Fähigkeiten sind auf ihre spezifischen Bedürfnisse abgestimmt (Miyashita et al., 1993). Relevante Konzepte müssen unterschieden werden können (Barlow, 1989; Zemel and Hinton, 1995; Edelman et al., 2002). Sie müssen zwischen bekannten Signalen einer Klasse so generalisieren können, dass Ausprägungsvarianzen dieser Klasse möglichst auch wieder richtig zugeordnet werden können. Das muss auch funktionieren, wenn Klassen aus Unterklassen zusammengesetzt sind (Biederman, 1995).

2.2.2 Effizienz

Biologische Individuen müssen um Ressourcen wie Nahrung und Fortpflanzungsmöglichkeiten kämpfen. Ihnen steht dafür nur begrenzt Energie, Zeit und Information über ihre Umgebung zur Verfügung. Da das visuelle Informationssystem den Individuen „eingebettet“ ist, muss es ebenfalls energie-

und zeiteffizient sein und mit unvollständigen Informationen auskommen können.

Die Evolution ist sehr herausfordernd in Bezug auf Reaktionszeit. Der Mensch benötigt typischerweise für eine Objekterkennung unter 150ms (Potter, 1976; Oram and Perrett, 1992; Thorpe et al., 1996).

Ein geringer Energieverbrauch ist ebenfalls herausfordernd. Olshausen and Field (1996) vermuten, dass die Repräsentationen im Gehirn deshalb „spärlich“ sind. Barlow (1961) argumentiert, dass laterale Inhibierung dafür genutzt werden könnte. Levy and Baxter (1996) vermuten, dass die Informationsverarbeitung aus Energieeffizienzgründen über Pulse (spikes) realisiert ist.

Geschwindigkeit und Energieeffizienz fordern in dieser Hinsicht optimierte Verarbeitungsarchitekturen. Solche Optimierungen sind in künstlichen visuellen Systemen erstmal nicht notwendig. Möglicherweise kann man also die unterliegenden Verarbeitungsprinzipien anschaulicher darstellen und implementieren.

Ein grundlegendes Prinzip von visueller Informationsverarbeitung ist die Reduktion von hochdimensionalen Sensorsignalen (Attneave, 1954). Das Ergebnis sollte für die Einbindung in kognitive Prozesse des Individuums geeignet sein. Auch dies ist eine Effizienzforderung, die die Repräsentation der Informationen betrifft. Die benötigte Informationsmenge könnte z.B. durch „Minimum Entropy Coding“ (Barlow, 1989) oder „Minimum Description Length“ (Zemel and Hinton, 1995) der auswertenden Szenen bestimmt werden. Im folgenden Abschnitt wird die Effizienz von verschiedenen Repräsentationen detaillierter betrachtet.

2.2.3 Repräsentation

Bis heute herrscht Unklarheit darüber, wie in biologischen visuellen Systemen kodiert und damit repräsentiert wird. Typischerweise werden Untersuchungen so durchgeführt, dass die Aktivierung einzelner Zellen oder Zellregionen bei unterschiedlichen visuellen Stimuli gemessen werden. Für eine wissenschaftliche Auswertung sollte dabei eine Wiederholbarkeit der Ergebnisse gegeben sein, die schon bei Experimenten mit primitiven Lebewesen nicht einfach zu erreichen ist. Es kann z.B. eine Rolle spielen, welche Stimuli vorher gezeigt wurden oder ob das Individuum mit anderen kognitiven Prozessen abgelenkt ist.

Wahrnehmung funktioniert nicht statisch. Ein konstanter Stimulus führt immer zu einem Rauschen der Aktivität auf minimalem Niveau. Der Wahrnehmungsprozess selbst benötigt Entropie der Sensorsignale und ist selbst auch immer nur im Übergang. Der wissenschaftliche Diskurs über visuelle

Repräsentationen dreht sich jedoch selbst bei Bewegtbilderanalyse hauptsächlich um das Erstellen von statischen strukturerhaltenden Zuordnungen (Homomorphismen) zu Sensorsignalen (Shepard, 1968; Shepard and Chipman, 1970; Suppes et al., 1994; Edelman and Duvdevani-Bar, 1997b; Palmer, 1999; Choe, 2002).

Objekt- oder betrachterorientierte Repräsentation

Die wahrscheinlich umfangreichste Diskussion in der wissenschaftlichen Gemeinschaft zum Thema visueller Repräsentationen dreht sich darum, ob diese objekt- oder betrachterorientiert realisiert sind.

Neurophysiologische und -psychologische Untersuchungen zeigen, dass biologische visuelle Systeme bestimmte Ansichten von dreidimensionalen Objekten bevorzugen. Sie generalisieren zwischen diesen Ansichten relativ schlecht (Cerella, 1986; Rock and DiVita, 1987; Bülthoff and Edelman, 1992; Edelman and Bülthoff, 1992; Logothetis et al., 1994, 1995; Bülthoff et al., 1995; Tarr et al., 1998; Booth and Rolls, 1998). Ullman (1989, 1995, 1999) schlägt ein Funktionsprinzip vor, nachdem das am besten passende gespeicherte Modell an den Eingabesignalen neu ausgerichtet wird. Zweidimensionale Struktur wird dabei nur zwischen betrachterunabhängigen lokalen Bildeigenschaften ausgewertet. Eine Reihe dieser Versuche wurde mit dreidimensional verbogenen Büroklammern durchgeführt. Das sind möglicherweise Objekte, für die das Gehirn nicht besonders gut dreidimensionale Repräsentationen und daher auch Tiefenrotationen vornehmen kann. Weil Objekte dieser Art im Alltag keine Rolle spielen, werden sie möglicherweise eher zweidimensional repräsentiert. Eine andere wichtige Kritik hat Hummel and Biederman (1992) angebracht. Die betrachterorientierte Repräsentation ist nur eine zweidimensionale Bildrepräsentation, die eine Semantik von Bildkomponenten vernachlässigt. Kleine Unterschiede im Bild können große Bedeutungsunterschiede für das menschliche Verständnis bedeuten und umgekehrt. Biederman and Cooper (1991) haben Experimente durchgeführt, die bevorzugte Bahnung der visuellen Verarbeitung durch vorab wahrgenommene Bilder (Priming) untersuchen und dabei festgestellt, dass für fragmentierte Ansichten das Priming funktionierte, selbst wenn hinterher nur das Komplement des Fragments gezeigt wurde. Priming und damit ein Teil Objekterkennung funktioniert also auf einer konzeptuellen Ebene. Hummel (2000) fügt an, dass betrachterorientierte Erkennung nicht möglich ist, wenn Objekte eine flexible Anzahl und unterschiedliche Anordnungen von Teilkonzepten haben, wie z.B. bei Fenstern oder Fahrzeugen. Tarr et al. (1998); Edelman and Newell (1998) halten dagegen, dass dies durch Kombinationen von kleineren betrachterorientierten Repräsentationen möglich

wäre. Es bleibt jedoch offen, wie eine Kombination erkannt werden soll, wenn sie in verschiedenen Anordnungen erscheint.

Marr and Nishihara (1978) schlugen eine objektorientierte Repräsentation vor, die anhand der Erkennung von Teilobjekten möglich wird. Zunächst werden Kombinationen von einfachen Bildelementen gruppiert und dann einfachen dreidimensionalen Repräsentationen zugeordnet. Diese werden in eine 2,5 dimensionale Skizze eingebaut, die die Orientierung der Oberflächennormalen enthält (Marr, 1982). In diese Skizze werden dann dreidimensionale Primitive, wie Zylinder, Quader oder Prismen (sog. Geons) eingefügt. Darauf basierende Objekterkennungssysteme Grimson (1989); Pentland (1990); Blanz and Vetter (2003) können solche Primitive in einfachen Szenen erkennen. Sobald jedoch allgemeine dreidimensionale Formen zugelassen werden, funktionieren sie kaum noch. Ein ambitioniertes Objekterkennungssystem ist „recognition by components“ (RBC) von Biederman (1987); Hummel and Biederman (1992); Biederman (1999). Dabei werden Geons und ihre räumliche Anordnung in Abhänigkeit von gruppierten Bildelementen im Eingabebild parametriesiert. Dabei werden virtuell Bindungen über zeitlich korrelierende Signale hergestellt. Obwohl das Verfahren aufwändig biologisch motiviert ist, können nur einfache Geone in Strichzeichnungen erkannt werden.

Wahrscheinlich ist in biologischen visuellen Systemen eine Mischung aus objekt- und betrachterorientierten Repräsentation gegeben. In höheren Verarbeitungsbereichen des Gehirns findet man jedoch hauptsächlich objektorientierte Repräsetationen Perrett et al. (1991); Logothetis et al. (1994). Die Repräsentationen sind verteilt, beschreiben auch Parametrisierungen (Wang et al., 1998) und hängen davon ab, was vorher an Kompositionen und Geometrie gelernt worden ist (Booth and Rolls, 1998). Parametrisierungen beschreiben z.B. die „Rundheit“ einer Kurve, oder die „Steilheit“ eines Winkels (Biederman and Bar, 1999; Vogels et al., 2001; Kayaert et al., 2003). Die Repräsentationen werden dabei invarianter gegenüber klassischen Varianzen (Vogels and Biederman, 2002).

Teilrepräsenationen

Ein Großteil aktueller Verfahren zum maschinellen Objekterkennen funktioniert in zwei Stufen. Die erste Stufe extrahiert Bildmerkmale. Diese spannen einen Merkmalsraum auf. Dieser ist in Bereiche unterteilt, die jeweils einem Konzept zugeordnet sind. Die zweite Stufe nimmt nun eine Klassifikation/Detektion vor, indem sie einen projizierten Punkt einem Bereich und damit einem Konzept zuordnet. Befürworter der betrachterorientierten Re-

präsentation können sich eine solche Realisierung auch in der biologischen Erkennung vorstellen (Edelman and Duvdevani-Bar, 1997a).

Einfache Bildmerkmale zeigen jedoch in der Regel keine statistische Relevanz in Bezug auf Konzeptunterschiede von Realweltgegenständen (Ruderman, 1994). Daher projizieren die Ansichten abstrakter Konzeptausprägungen meist sehr verstreut in den Merkmalsraum. Zum Beispiel würde die Abbildung eines Vogels durch typische Merkmalsextraktionsstufen völlig unterschiedliche Merkmale extrahieren - je nachdem ob er fliegt oder sitzt. Eine korrekte und gut generalisierende Klassifikation ist bei klassischen und abstrakten Varianzen mit zwei Stufen kaum möglich.

Das biologische Vorbild realisiert offenbar nicht nur eine Extraktions- und eine Detektionsstufe. Verarbeitung und Repräsentation von Bildinformation erfolgt entlang eines Pfades mit sehr vielen Stufen (Van Essen and Maunsell, 1983). Entlang des Pfades beschreiben die Repräsentationen der Stufen immer komplexere (Teil-)Bildinhalte (Oram and Perrett, 1994; Van Essen and Drury, 1997). Zusätzlich findet man, dass die Repräsentationen immer invarianter auf Ausprägungsvarianzen einer Konzeptabbildung regieren - sie also spezifischer auf das abstrakte Konzept werden (Tanaka et al., 1991; Kobatake and Tanaka, 1994). Abbildungsvarianzen von Konzepten sind im Allgemeinen sehr komplex. Um sie detektieren und repräsentieren zu können, werden offenbar effiziente Teilrepräsentationen eingesetzt, die flexibel zu einer Szenenbeschreibung kombiniert werden können (Perrett and Oram, 1993; Oram and Perrett, 1994). Teilrepräsentationen müssen so zusammengesetzt werden, dass typische Varianzen nicht das Ergebnis einer Detektion verändern.

Die Anzahl der zu verwendenden Teilrepräsentationen richtet sich nach der Komplexität der Objekte und danach wie eine Unterscheidbarkeit von Teilobjekten für das betrachtende Lebewesen von Bedeutung ist (Palmer, 1977). Aus Effizenzgründen werden wahrscheinlich zur Repräsentation von unterschiedlichen Konzepten möglichst viele gemeinsam nutzbare Teilkonzepte verwendet. Wenn es im Inferior-Temporalen-Cortex eine große verteilte Aktivität während einer Gesichterkennung gibt, dann bedeutet das nicht, dass eine Menge von Neuronen selektiv für Gesichter sind. Es ist vielmehr davon auszugehen, dass diese aktiven Neurone passende Teilrepräsentationen zur Repräsentation des betrachteten Gesichtes sind. Sie tragen detaillierte Gesichtscharakteristik bei und damit auch Unterscheidungsmöglichkeiten Perrett et al. (1998).

Kodierung

Wenn man davon ausgeht, dass elektrische Impulse der Nervenzellen Informationen repräsentieren, bleibt immer noch offen, wie diese kodiert sind. Frühe Einschätzungen betrachteten jede Zelle als einem bestimmten Konzept zugeordnet (Barlow (1972), sog. „Großmutterzellen“). Neuere Untersuchungen zeigen, dass Information eher regional verteilt repräsentiert ist. (Oram and Perrett, 1994; Gochin et al., 1994; Földiák and Young, 1995; Gauthier et al., 2000; Tsunoda et al., 2001).

Die Kodierung von Information auf der Retina und den ersten Folgestufen entlang des Verarbeitungspfades kann noch relativ gut bestimmt werden. Dort werden insbesondere natürliche Realweltszenen effizient und spärlich kodiert (Olshausen and Field, 1996; Bell and Sejnowski, 1997; Hyvärinen et al., 2009). In höheren Verarbeitungsstufen sind Repräsentationen in verteilten Aktivierungen gegeben. Die Verteilung ändert sich mit Varianzen (Wang et al., 1998). Wie dabei kodiert wird, ist noch weitgehend unbekannt.

Ebenso unklar ist die Kodierung von zwei- und dreidimensionalen Anordnungen (Peters, 2000). Dies sind jedoch wesentliche Eigenschaften, die auch jedes künstliche Bildanalysesystem realisieren muss.

Für die Informationskodierung spielen wahrscheinlich auch zeitliche Aspekte der Nervenimpulse eine Rolle. Darauf wird im nächsten Abschnitt noch eingegangen. Maass and Bishop (1999) geben einen Überblick über mögliche Realisierungen zeitabhängiger Kodierungen.

2.2.4 Integrieren von Repräsentation

Auf dem visuellen Verarbeitungspfad müssen Teilrepräsentationen erzeugt werden. Dazu wird auf vor- und nachgeschalteten Schichten und auch in der betrachteten Schicht Information ausgewertet. Dieser Abschnitt beschäftigt sich damit, wie diese Information zusammengetragen und ausgewertet wird.

Beim Integrieren von Repräsentationen werden neue Repräsentationen mit komplexerer Spezifität erzeugt. Dabei geht typischerweise Information über Details verloren. Es gibt jedoch auch die Meinung, dass bei der biologischen Verarbeitung keine Information verloren ginge und der Wahrnehmungsvorgang umkehrbar sei (Hinton and Ghahramani, 1997).

Kontext und Integration

Man findet starke Kontextabhägigkeiten entlang des gesamten Verarbeitungspfades (Biederman et al., 1973; Bar and Shimon, 1993). Dabei be-

ginnt die Kontextabhängigkeit in der direkten Umgebung von Sehzellen der Netzhaut, die ihre direkt Umgebung auf Hell-Dunkel-Kontraste untersuchen und wird dann in höheren Verarbeitungsbereichen räumlich weitläufiger. Die Abhängigkeiten werden dabei abstrakter. Es ist z.B. wesentlich, wie wahrscheinlich ein zu detektierendes Objekt in der Szene a priori gegeben ist. Der Erkennungsvorgang für ein Sofa dauert z.B. länger, wenn es in einer Straßenszene abgebildet ist, als wenn es in einer Wohnraumszene abgebildet ist (Biederman, 1981). Es spielt auch eine Rolle, ob das zu erkennende Objekt eine typische Ausrichtung und Größe hat (Biederman et al., 1973).

Kontextabhängigkeit wird nicht mit starren Schablonen analysiert. Sie kann auch funktional sein. So kann es z.B. beim Erkennen eine Rolle spielen, ob ein umgebendes Merkmal gleichartig ist. Wenn beispielweise ein Fensterrahmen grün ist, wird das dazu beitragen, umliegende Fenster schneller wahrzunehmen, wenn sie ebenfalls so ausgeprägt sind. Biologische Objekterkennung kann bestimmte kontextuale Anordnungen erkennen, die im Sinn von Repräsentation effizient sind (Heitger et al., 1998; Fiser and Aslin, 2001). Dabei spielen Konstanzmechanismen, wie z.B. Farbgleichheit oder Colinearität eine Rolle (Ross and Mingolla, 1998). Es kann aber auch von der Gestalt der Elemente selbst abhängen (Humphreys et al., 1998). Dieses „Gruppieren“ von Bildteil-Repräsentationen wird wahrscheinlich auch zum mentalen vervollständigen von Objektteil-Ansichten verwendet. Es wird wahrscheinlich auch auf vielen Bereichen entlang des Verarbeitungspfades eingesetzt (Selinger and Nelson, 1999). Grossberg and McLoughlin (1997); Ross et al. (2000) versuchen Gruppierungseffekte mit der Funktion von Hypercomplexen Zellen zu erklären. Dies gelingt aber nur schwer - ähnlich wie eine Funktionsanalyse für Symmetrie (Oka et al., 2001; van Tonder and Ejima, 2000) oder Linienvervollständigung (Chen and Lin, 1998; Li, 1998). Es gibt noch eine Reihe anderer präferierter kontextualer Anordnungen (Sarkar and Boyer, 1993).

Binden

Objektdetektoren reagieren spezifisch auf bestimmte Eigenschaftskombinationen von Objektabbildungen. Oft sind aber mehrere Objekte im Bild gegeben und dann ist fraglich, welche Eigenschaft zu welchem Objekt gehören soll. Wenn z.B. die Eigenschaften „dreieckig“, „quadratisch“, „rot“ und „grün“ global detektiert werden können und es ein rotes Quadrat und ein grünes Dreieck im Bild gibt, dann ist aus den Ergebnissen nicht mehr zu bestimmen, welche der detektierten vier Eigenschaften tatsächlich im Eingabebild in Kombination gegeben sind. Dies ist als das sogenannte Bindeproblem bekannt (von der Malsburg, 1994, Reprint from 1981 sowie Milner, 1974;

Feldman and Ballard, 1982). Wenn man statt der globalen Eigenschaftsdetektoren lokale benutzt, hat man das Problem, dass für alle Bildpositionen und Eigenschaftkombinationen ein einzelner Detektor bereitgestellt werden muss. Der Aufwand ist exponential in der Anzahl der Eigenschaften und ist daher praktisch nicht für Realweltobjekte anwendbar.

Wie die biologische Bildanalyse das Bindeproblem löst, ist bisher unbekannt. Eigenschaftkombinationen werden dort sogar über Verarbeitungspfade gebunden, die offenbar getrennt sind („wo-“ und „was-“ Pfade Ungerleider and Mishkin, 1982; Merigan and Maunsell, 1993). Sogar unterschiedliche Sinnesmodalitäten können Inhalte aneinander binden (von der Malsburg and Schneider, 1986).

Um den exponentiellen Aufwand zu umgehen, schlug Wickelgren (1969) vor, mehrere Schichten mit Teildetektoren einzuführen. Dabei wird in jeder Schicht nur ein wenig Information über die Position verloren, so dass ein Zuordnen der Eigenschaften zu den Bildregionen noch möglich ist. Die Idee wurde evaluiert (McClelland and Rumelhart, 1981) und weiterentwickelt (Mozer, 1991; Mel and Fiser, 2000). Sie ist biologisch motiviert (Hubel and Wiesel, 1962, 1965; Grossberg, 2003) und findet als sog. „Pooling“ Verwendung in einigen hierarchischen Verfahren zur Objekterkennung (z.B.Fukushima et al., 1983; LeCun et al., 1990; Teichert and Malaka, 2002, 2003).

Lange Zeit wurde die Aktivität von Nervenzellen als eine gemittelte „Feuerrate“ von Impulsen modelliert. Als Gray et al. (1989) und Eckhorn et al. (1988) herausgefunden haben, dass die zeitlichen Beziehungen zwischen Neuronen bei der Informationsverarbeitung eine Rolle spielen und sich Zellverbände gegenseitig synchronisieren können, wurden zeitliche Aspekte der Impulse genauer untersucht. Z.B. könnten sie in unteren Schichten des Verarbeitungspfades zur Segmentierung genutzt werden (Engel et al., 1991a,b; Grossberg and Somers, 1991; Engel et al., 1992; Singer and Gray, 1995). Einfache Oszillatormodelle konnten das Verhalten nachahmen (Baldi and Meir, 1990; König and Schillen, 1991; Schillen and König, 1991; Wang et al., 1991; von der Malsburg and Buhman, 1992; Vorbrüggen and v.d. Malsburg, 1995; Ranganath and Kuntimad, 1996; Teichert and Malaka, 2000; Knoblauch and Palm, 2003; Nakano and Saito, 2004). Neben Segmentierungen, könnten über Synchronitäten auch unterschiedliche Sinnesmodalitäten in Bezug gesetzt und damit Eigenschaften gebunden werden (Atiya and Baldi, 1989; König and Engel, 1995; Vaadia et al., 1995; Oram and Perrett, 1996; Wehr and Laurent, 1996; Prut et al., 1998). Möglicherweise können Bildinhalte, die aus verschiedenen Blicksackaden in den visuellen Verarbeitungspfad eingeleitet wurden, auch über Synchronisation in Bezug gesetzt werden (Grossberg and Somers, 1991; Parker and Newsome, 1998).

Allerdings gibt es auch Hinweise, dass temporale Aspekte zwar lokal eine Rolle spielen mögen - überregionale Zellverbindungen aber eher die Aktivitäten aufintegrieren (Brody, 1998; Oram et al., 1999, 2001; Lücke, 2002). Es gibt verschiedene Möglichkeiten, wie Informationen in zeitlichem Verhalten der Nervenimpulse kodiert sein kann (Panzeri et al., 1999; Oram et al., 2002; Delorme, 2003; Wennekers and Ay, 2003).

Vorwärts und rückwärts gerichtete Verarbeitung

Ein Großteil der neuronalen Verbindungen sind entlang des visuellen Verarbeitungspfades rückwärts gerichtet oder wirken innerhalb einer Schicht auf regional benachbarte Zellen (Van Essen and Maunsell, 1983; Felleman and Van Essen, 1991; Braitenberg and Schüz, 1998).

Es wird angenommen, dass die vorwärts gerichteten Signale invariant detektierbare Signalteile liefern. Diese sind eher Vorschläge, was an Inhalt im Bild gegeben sein kann. Auf der höheren Schicht wird dann assoziiert. Das sich einstellende Muster hängt von den vorwärts gerichteten Signalen und sehr stark von den Aktivierungen in der Nachbarschaft ab (Mumford, 1992; Rao and Ballard, 1997). Die Aktivitäten in solchen autoassoziativen Netzwerken konvergieren in Muster, die durch Lernen der Verbindungsstärke bestimmt werden (Hopfield, 1982; Cohen and Grossberg, 1983; Kosko, 1988). Die Muster könnten so eingestellt sein, dass sie nur konsistente Repräsentationen zulassen und durch inhibierende Rückwirkung die Aktivitäten vorgeschalteter Schichten disambiguieren (Mumford, 1992, 1994; Connor et al., 1997; Hinton and Ghahramani, 1997; Liang and Wang, 2003). Modelle dieser Funktion stützen die Annahme (Rao and Ballard, 1996, 1997; Batlle et al., 2000; Heinke and Humphreys, 2003).

Bei der Verarbeitung innerhalb einer Schicht spielen neben Kontextabhängigkeiten Gruppierungseffekte eine Rolle (Arbib, 1995). Höhere Schichten haben auch bereits Funktionen eines Kurzzeitgedächtnisses (Miyashita and Chang, 1988). Dies kann Teil eines Aufmerksamkeitssystems sein, dass die Erkennung moduliert (Desimone and Duncan, 1995; Luck et al., 1997). Auch dazu gibt es Modellierungen, die allerdings noch keine generelle Lösung für das Huhn-Ei-Problem „Objektsuche und Objektklassifikation“ bieten (Salinas and Abbot, 1997; Hamker, 1999; Hamker and Worcester, 2002; Olshausen et al., 1993; Heinke et al., 2002).

Von der Aufmerksamkeitsforschung gibt es einen fließenden Übergang zur Bewußtseinsforschung (Crick and Koch, 1995; Logothetis, 1998; Crick and Koch, 1998; Kanwisher, 2001; Tong, 2003). Modellierungen, die versuchen Bewußtsein zu implementieren, konnten bisher nicht für Bilderkennungssysteme genutzt werden (Taylor, 1997).

2.2.5 Lernen

Lernen in biologischen Bilderkennungssystemen stellt Spezifitäten von Teilrepräsentationen so ein, dass biologisch relevante Stimuli angemessen unterschieden werden können (Miyashita et al., 1993). Dabei ist eine hohe Flexibilität gegeben, so dass eine Adaption an unterschiedliche visuelle Umgebungen möglich ist. Je höher die Teilrepräsentationen im Verarbeitungspfad liegen, desto mehr sind die Teilrepräsentationen von der visuellen Historie eines Individuums abhängig.

Es spielen unterschiedliche Zeitskalen eine Rolle. Wahrnehmung kann sich im Minutenbereich auf eine bestimmte Aufgabe einstellen (Priming) oder auch einmal gesehenes ein Leben lang speichern (One-Shot-Learning). Meist ist beim visuellen Lernen ein Konzept zugeordnet (überwachtes Lernen). Gerade auf unteren Schichten findet aber auch eine Selbstorganisation statt (unüberwachtes Lernen). Überwachtes und Unüberwachtes Lernen können zusammenwirken (Grossberg, 1980).

Es gibt zahlreiche Erkenntnisse über biochemische Wirkungsweisen beim Lernen. Bisher fällt es jedoch schwer, die dahinter liegenden allgemeinen Prinzipien aufzudecken (Kandel et al., 2000). Es können jedoch einige Teilprinzipien ausgedrückt werden: Korrelierende Aktivitäten führen zur Verstärkung der zugehörigen Verbindung (Hebb, 1949). Es wird versucht, Redundanz zu reduzieren (Barlow, 1989; Redlich, 1993; Zemel and Hinton, 1995) oder Entropie zu maximieren (Linsker, 1989, 1992). Es gibt einen Wettbewerb, wobei sich nur die passensten Spezifitäten weiter verändern (von der Malsburg, 1990). Ein Vorschlag für ein mögliches Teilprinzip ist, dass sich relevante Signalanteile relativ langsam gegenüber varianten Signalanteilen verändern (Földiák, 1991; Becker, 1993; Stone, 1996). Auch dies ist biologisch motiviert und wird aktuell als „Slow Feature Analysis“ modelliert (Wiskott and Sejnowski, 2002; Berkes and Wiskott, 2003; Wiskott, 2003).

Ein Problem ist der „Fluch der Dimensionen“, der für Neuronale Netzwerke bedeutet, dass die Anzahl der benötigten Trainingsmuster exponential zur Anzahl der verwendeten Neuronen ist (von der Malsburg, 1995). Man kann jedoch die Musteranzahl reduzieren, wenn man nicht die optimale Funktion finden muss (Vapnik, 1999), andere Neuronenmodelle verwendet (Geman et al., 1992) oder zusätzliche Randbedingungen einführt (Poggio et al., 1985). Generell gilt, dass die kleinstmögliche Flexibilität die besten Ergebnisse ergibt (Occams Razor, z.B. in Duda et al., 2002, S.465).

2.2.6 Varianzen

Beim Objekterkennen bedeutet Spezifität, dass nur bestimmte Bildinhalte erkannt werden - ein Objekterkennungssystem reagiert spezifisch auf diese Bildinhalte und diskriminiert andere. Z.B. sollen Autos erkannt werden und Häuser nicht. Dahingegen bedeutet Invarianz, dass bestimmte Bildinhalte in verschiedenen Ausprägungen (Varianzen) gegeben sein dürfen und dies die Erkennung nicht beeinträchtigt. Für eine Gebäudeerkennung soll es z.B. keine Rolle spielen, welche Ausprägung der Himmel hat. Der Gebäuderkenner soll invariant gegenüber diesen verschiedenen Hintergrundausprägungen sein.

Für Objekterkennungssyteme, die mit mehreren Stufen und Teilrepräsentationen realisiert sind, wurde bisher Invarianz meist so interpretiert, dass alle Informationen, die zur Detektion eines Teilkonzeptes beigetragen haben, für Folgestufen nicht mehr sichtbar sein müssen. Mit Varianzen dieser Signale muss nur in der Detektionsstufe umgegangen werden. Folgende Detektionsstufen bekommen nur noch die Information, dass das Teilkonzept gegeben ist - nicht aber dessen Ausprägung. Ob z.B. eine Uhr mit Zeigern oder Digitalanzeige ausgestattet ist, muss für das Konzept „Uhr“ keine Rolle spielen. Diese Varianz kann auf der Stufe zur Erkennung der Gehäuse Frontansicht aufgelöst d.h. invariant gemacht werden. Aus dieser Interpretation wurden ausgehend von Pitts and McCullough (1947) etliche biologisch motivierte Verfahren zum Bildverstehen erstellt. Sie verwenden eine Hierarchie von Teilkonzepten, die stets nur invariante Signale zur nächsthöheren Detektionsstufe weiterleiten (Fukushima, 1975; Biederman, 1987; LeCun et al., 1990; Olshausen et al., 1993; Wallis and Rolls, 1997).

Das Erkennen und explizierte Repräsentieren von varianten Signalanteilen ist jedoch von großer Bedeutung für das Bildverstehen. Zum Beispiel ist es häufig wichtig zu wissen, ob die Ausprägung eines Teilkonzeptes innerhalb eines bestimmten Bereiches gegeben ist. Beim Erkennen einer Uhr ist es wesentlich, die Information zu erhalten, dass entweder Zeiger *oder* eine Digitalanzeige in der Frontansicht gegeben sind. Sonst wäre eine Armbanduhr beispielsweise nicht von einem Armreif zu unterscheiden. Generell kann die explizite Repräsentation von Varianzen für benachbarte Bildregionen hilfreich sein, wenn dort keine eindeutige Detektion möglich ist. Dies ist für Varianzen der Fall, die typischerweise regionale Ausbreitungen haben, wie z.B. perspektivische Verzerrungen, Texturausprägungen oder Stilausprägungen. Schließlich kann es wesentlich sein, eine bestimmte Ausprägung eines vorgeschalteten Teilkonzeptes zu erkennen. Wenn z.B. die Konzepte Digitaluhr und Analoguhr unterschieden werden sollen, muss sowohl das übergeordnete Konzept „Uhr“ als auch die Teilkonzepte „analoges Zifferblatt“ und

„digitales Zifferblatt“ erkannt werden können.

Die Detektion eines Teilkonzeptes kann sowohl von invarianten, als auch von varianten Signalen abhängen. Für ein Verfahren zum Erkennen von Gegenständen mit starken Abbildungsvarianzen ist es sinnvoll, sowohl variante als auch invariante Signale explizit zu kodieren und zu verarbeiten. Das ist auch bei neu zu lernenden Konzepten wichtig. Dort wird oft ein zunächst invariantes Signal zu einer Spezifität. Wenn ein Objekterkenner z.B. die Konzepte „Pferd“ und „Esel“ erkennen kann, dann ist ein Signal, das Streifen eines „Zebras“ beschreibt, zunächst invariant zu verarbeiten. Wenn dieses Konzept jedoch auch erkannt werden soll, muss es mit anderen Signalen zu einer Spezifität zusammengefasst werden.

Für klassische Bilderkennungssysteme sind bisher viele Varianzen identifiziert, die in unterschiedlichen Objektausprägungen begründet sind: Position, Größe, Orientierung, Umriss, Projektion, Helligkeit, Farbe, Bewegung, Verdeckung, Rauschen (Rock, 1985; Palmer, 1999). Untersuchungen an der biologischen visuellen Informationsverarbeitung lassen vermuten, dass Varianzen stets dort verarbeitet werden, wo Bildinhalte repräsentiert werden, die von der Varianz betroffen sind (Oram and Perrett, 1994). Also auf der Stufe der zugehörigen Teilrepräsentation. Die anatomischen Verarbeitungsstrukturen sind relativ homogen, so dass man davon ausgehen kann, dass die o.g. Varianzen nach einem gleichen Prinzip verarbeitet werden. Bisher ist es jedoch nicht gelungen ein solches Verarbeitungsprinzip aufzudecken und zu beschreiben.

Im Folgenden werden einige Varianzen und ihre Eigenheiten bei der biologischen Informationsbearbeitung beschrieben. Für Realweltszenen tritt meist eine Mischung auf.

Verschiebung und Skalierung

Biologische Bilderkennungssysteme erreichen Verschiebungs- und Skalierungsinvarianz entlang ihres Verarbeitungspfades. Frühe Stufen zeigen eine regionale Invarianz bezüglich ihrer Spezifität (Hubel and Wiesel, 1962). Höhere Stufen zeigen komplexere Repräsentationen, die über weite Teile des Sichtfeldes invariant bezüglich Verschiebung und Skalierung sind (Perrett et al., 1982; Tanaka, 1996). Psychologische Untersuchungen zeigen je nach Untersuchungsmethode homogene Verschiebungs- und Skalierungsinvarianz (Biederman and Cooper, 1991, 1992) oder aber bestimmte Präferenzen für bestimmte Objektgrößen und -positionen (Jolicoeur, 1987; Nazir and O'Regan, 1990; Dill and Fahle, 1997; Dill and Edelman, 1997).

Wie Verschiebungs- und Skalierungsinvarianzen realisiert werden, ist bisher unklar (Würtz, 1999; Wiskott, 2004). Es gibt Hinweise darauf, dass eine

rückwärtsgerichtete Beeinflussung des Verarbeitungspfades dabei eine Rolle spielen (Kosslyn et al., 1992; Hellige and Cumberland, 2001).

Im nächsten Kapitel werden Verfahren beschrieben, die unterschiedlichen Techniken, wie z.B. „Pooling“ einsetzen, um in gewissen Grenzen Verschiebungs- und Skalierunginvarianz zu erreichen. Solange aber für die Extraktion der Bildinformationen Filter mit festen Maskengrößen verwendet werden, ist Skalierungsinvarianz grundsätzlich problematisch, da zwischen den Ergebnissen schlecht generalisiert werden kann (Würtz, 1994; Lindeberg, 1994).

Rotation

Wie bei Verschiebungs- und Skalierungsinvarianz finden sich für Invarianzen gegenüber Rotation von Objekten einige Untersuchungen, die vermuten lassen, dass diese auch über weite Teile des biologischen Verarbeitungspfades gegeben sind (Biederman and Gerhardstein, 1993). Ein größerer Teil findet jedoch Hinweise für Spezifitäten für verschiedene Ansichten - also eher betrachterorientierte Repräsentationen (Warrington and Taylor, 1973; Humphrey, 1984; Tarr, 1995; Tanaka, 1997; Tarr et al., 1998; Ashbridge et al., 2000). Diese Eigenschaft nimmt zu, wenn die visuelle Struktur komplexer wird (Newell, 1998; Hayward and William, 2000). Bei Objektrotationen treten starke Veränderungen der visuellen Struktur auf. Das ist möglicherweise der Grund für separate Repräsentationen (Moses et al., 1994; Biederman, 2001). Die benötigte Zeit bis zur Erkennung erhöht sich mit steigendem Winkel gegenüber bekannten Prototypenansichten (Shepard and Metzler, 1971; Newell and Findlay, 1997). Generalisierungsleistungen verbessern sich, wenn mehrere Prototypenansichten bekannt sind (Peissig et al., 2002).

Abstrakte Varianzen

Zu diesem Typ von Varianzen sind bisher kaum Untersuchungen angestellt worden. Das liegt sicherlich auch daran, dass die Diskussion über die Verarbeitung von klassischen Varianzen schon kontrovers genug geführt wird.

Bemerkenswert sind jedoch die Darstellungen, die Oram and Perrett (1994) basierend auf Forschungsergebnissen von Tanaka et al. (1991) angefertigt haben. Abbildung 2.1 zeigt, welche Spezifitäten Kolumen von Neuronen in höheren Schichten des visuellen Verarbeitungspfades haben können. Die konkreten Ausprägungen der (Teil-)Abbildungen sind dabei teilweise sehr unterschiedlich. Invariant ist meist eine Anordnung von Teilmustern, die selbst eine hohe Ausprägungsvarianz haben dürfen. Die Abbildung zeigt, wie die Spezifitäten genutzt werden können, um unterschiedliche Objekte

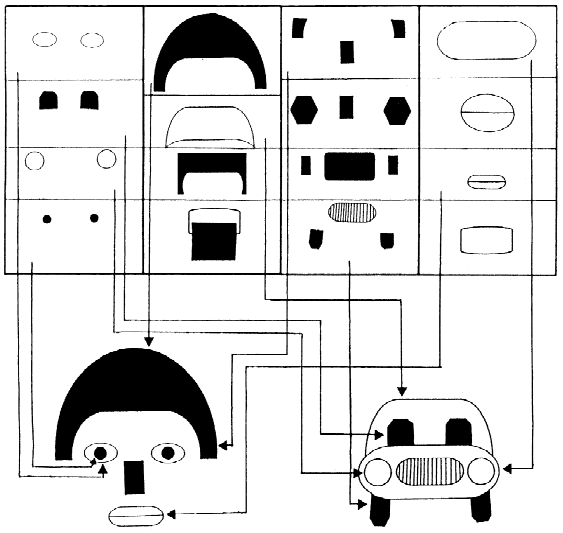


Abbildung 2.1:
Zellselektivitäten von vier Neuronenkolumnen (Spalten) in höheren Schichten des visuellen Verarbeitungspfades (aus Oram and Perrett, 1994). Die Zeilen zeigen für welche Teilabbildungen Neurone einer Kolumne spezifisch sein können. Das sind meist invariante Anordnungen von Teilabbildungen, die selbst eine hohe Ausprägungsvarianz zeigen dürfen. Mit diesen Selektivitäten lassen sich unterschiedliche Objekte repräsentieren.

zu repräsentieren. Für das Autobeispiel würden mit diesem Neuronensatz auch eine große Menge von abstrakten Varianzen erkannt werden können.

Welche (Teil-)Spezifitäten gegeben sind, ist in jedem Lebewesen unterschiedlich und liegt sicherlich an der visuellen Lernhistorie des Individuums und seinem Diskriminierungsbedarf der Szenenobjekte. Wenn Szenenobjekte abstrakten Varianzen unterliegen, müssen unterschiedliche Teilrepräsentationen für eine Erkennung zugelassen werden. Das dies möglich ist, zeigen die Untersuchungen von Tanaka et al. (1991); Tanaka (1996).

Mit diesen Prinzipien wären auch topologische Varianzen zu verarbeiten, die hier als eine spezielle Unterart von abstrakten Varianzen betrachtet werden.

2.3 Diskussion und Zusammenfassung

Es gibt keine klassische Methodik, um die in Abschnitt 1.2.4 beschriebene Fragestellung zu beantworten und die Aufgabe zu erfüllen, ein entsprechendes Erkennungsverfahren zu erstellen. Die gewünschten Erkennungsleistungen werden nur von biologischen visuellen Systemen erreicht. Bisher ist es nicht gelungen, deren zugrundeliegenden Verarbeitungsprinzipien aufzudecken. Aus heutiger Sicht bleiben viele wichtige Aspekte unklar, wie sie auch schon als Herausforderungen in Abschnitt 2.1 beschrieben wurden. Jedes Verfahren, welches universelles Objekterkennen realisieren soll, muss

eine Lösung für diese Herausforderungen finden. Dafür muss die Biologie nicht komplett nachgebaut werden. Aber ohne ein tiefergehendes Verständnis biologischer Funktionen, kann auch nicht identifiziert werden, welche Gegebenheiten weggelassen werden könnten und ob Funktionen, wie ein z.B. ein Aufmerksamkeitssystem überhaupt separierbar wären.

Die Anforderungen von technischen und biologischen Bilderkennungssystemen sind (bisher) grundlegend verschieden. Während technische Systeme meist eine gut abgrenzbare Aufgabe erfüllen sollen, sind biologische Bilderkennungssysteme in komplexe Umgebungen und Anforderungen eingebettet. Möglicherweise genügt daher auch eine einfachere „Implementierung“ biologischer Verarbeitungsprinzipien.

Vielleicht gibt es auch noch effizientere Implementierungsbausteine, als die, die der Biologie zur Verfügung stehen. So wird in der Biologie eine relativ starr dreidimensional verteilte Struktur aus Zellen und Nervenbahnen verwendet, um Repräsentationen zu verarbeiten. Eine andere Architektur könnte flexible Elemente zum Repräsentieren verwenden. Dafür bräuchte man möglicherweise keine komplexen zeitlichen Kodierungen mehr. Solche Lösungen wird man nur finden, indem man ausprobiert - Verfahren mit aussichtsreichen Eigenschaften kombiniert, neu erstellt und aus den Ergebnissen lernt, welche Eigenschaften eine wichtige Rolle spielen können.

Kapitel 3

Verfahren zur Bilderkennung

Welche Verfahren werden betrachtet

Im vorigen Kapitel wurden schon einige Verfahren zum Bildverstehen genannt. Diese wurden meist erstellt, um bestimmte Hypothesen über die biologische Funktionsweise zu überprüfen. Für den Einsatz in Realweltanwendungen fehlen ihnen wichtige Fähigkeiten. In diesem Abschnitt sollen Verfahren betrachtet werden, die dafür ausreichend Funktionalität mitbringen. Dabei besitzen sie immer noch Einschränkungen in der Verwendbarkeit. Beispielsweise werden bestimmte Hintergründe vorrausgesetzt oder es sind nur bestimmte Bildstrukturen detektierbar. Meist sind die Varianzen, unter denen Objektabbildungen detektierbar bleiben, beschränkt. Der Übergang von den Verfahren aus Kapitel 2 zu den hier betrachteten ist fließend.

Oft sind aus interessanten Verfahren, viele Weiterentwicklungen entstanden, deren Eigenschaften ähnlich bleiben. Die Folgenden Betrachtungen beziehen sich dann auf typische Vertreter.

Bewertung der Verfahren

In Abschnitt 2.1 wurden schon Herausforderungen ausgedrückt, die für ein Bilderkennungssystem wesentlich sind. Die Bewertung der Eigenschaften der existierenden Verfahren soll auch nach dieser Gliederung erfolgen.

Zunächst soll es darum gehen, wie ein funktionierendes Verfahren überhaupt erstellt werden kann. Die Fragen nach der Effizienz werden daher nicht betrachtet.

Für die restlichen Gliederungspunkte ergeben sich noch einige Erläuterungen:

Repräsentation

Wie wird in den künstlichen Systemen Information repräsentiert?

- Was wird repräsentiert?

  Aus Bildern wird Information extrahiert. Die Ergebnisse werden als Zwischen- oder Endergebnis repräsentiert. Was steckt an Information in den Ergebnissen und was wird verworfen und ist dann nicht mehr verfügbar?

- Wie werden Strukturen und Anordnungen kodiert?

  Zeitgenössische Rechensysteme bieten gute Unterstützung für Bildrepräsentationen als zweidimensionales Raster. Zur Auswertung von Bildstrukturen werden daher meist Rasterrepräsentationen verwendet. Dadurch ergeben sich jedoch Diskretisierungsprobleme, wenn Bildinformation durch klassische Varianzen verschoben wird.

  Bei maskenbasierten (Filter-)Auswertungen ist Information über die Anordnung von Bildstrukturen intrinsisch enthalten. So müsste es aber für jede Anordnung einen extra Filter geben. Effizienter sind Repräsentationen, die Eigenschaften von Anordnungen parametrisieren. Wie kann das erreicht werden?

- Sind innere Repräsentationen aussagekräftig?

  Was kann aus den erstellten Repräsentationen noch an weiterer Information abgeleitet werden? Wenn z.B. ein Objekt detektiert wurde, können dann noch Aussagen über seine Ausprägung und Anordnung in der Szene gemacht werden? Kann ein detektiertes Objekt auch im Eingabebild segmentiert werden? Ermöglichen interne Repräsentationen ein Benennen und Segmentieren von Teilobjekten?

Integration von Repräsentation

Verfahren zum Bildverstehen müssen überprüfen, ob bestimmte Kombinationen von Bildstrukturen gegeben sind. Dazu müssen umgebende Repräsentation zusammengeführt und ausgewertet werden. Wie ist das realisiert?

Bei der Bewertung der Verfahren soll unter dem Begriff „Integration“ stets dieses Zusammenfassen und Auswerten verstanden werden.

- Wie werden Repräsentationen integriert?

  Welche Vorraussetzungen müssen erfüllt sein, um neue Repräsentationen zu erstellen? Für welche Region ist das Ergebnis aussagekräftig?

- Wie wird Kontext ausgewertet?

  In welchem Bereich wird umgebende Bildinformation ausgewertet? „lokal“ bedeutet nur in der direkten Nachbarschaft. „regional“ beschreibt einen größeren Bereich um die auszuwertende Bildposition - typischerweise den Bereich einer Filtermaske. „überregional“ beschreibt Anordnungen, die nicht mit einer einzelnen Auswertungsschicht in Bezug zu setzen sind. Schließlich beschreibt „global“ den gesamten Bildbereich.

  Wenn es sich um Verfahren mit mehreren Schichten handelt, ist interessant, wie der Kontext dort ausgewertet wird. Dort können auch andere Metriken verwendet werden, als in der Bildebene.

- Wie werden Eigenschaften an Repräsentationen gebunden?

  Können mit den Repräsentationen noch Attribute aus dem Detektionsvorgang beschrieben werden, z.B. Farben oder andere Ausprägungen? Wie wird das realisiert? Wie sind diese Informationen später auszudrücken?

  Sofern dieser Punkt nicht erwähnt wird, findet kein Binden statt.

- Wie fließt bekanntes Wissen in den Erkennungsvorgang ein?

  Informationen aus dem Eingabebild werden mit bereits erlernten Informationen verglichen. Der Vergleich kann schon das Wissen enthalten, wie z.B. bei einem globalen Histogramm über Bildeigenschaften oder der Gestalt von Filtermasken. Um mit abstrakten Varianzen umgehen zu können, werden jedoch komplexere Modelle von bekanntem Wissen verwendet. Wie sind diese repräsentiert und wie wirken sie im Erkennungsvorgang?

Lernen

- Wie funktioniert Lernen in Objekterkennungssystemen?

  Wie werden die Repräsentationen von bekanntem Wissen erstellt?

Varianzen

Wie gut ein Objekterkennungssystem mit Varianzen umgehen kann ist ein wesentliches Merkmal. Im Folgenden wird der Begriff „Auffangen“ für eine invariante Erkennungsleistung bezüglich eines Varianztyps verwendet.

- Wie funktioniert Erkennen unter klassischen Varianzen?

Unter klassischen Varianzen soll in diesem Beitrag die Veränderung von Bildinhalten durch Beleuchtung, Verschiebung, Skalierung, Rotation, perspektivischer Transformation oder topologieerhaltende Verzerrung verstanden werden.

Varianzen, die nicht topologieerhaltent sind, zählen nicht mehr zu den klassischen Varianzen, z.B. unterschiedliche Fensteranzahl und Anordnung von Fenstern in Gebäudeansichten.

- Wie funktioniert Erkennen unter abstrakten Varianzen?

  Wie schaffen es Bilderkennungsverfahren, abstrakte Varianzen aufzufangen? Wie komplex dürfen die Objekte dabei sein?

  Der Begriff Ausprägungsvarianzen soll topologische und abstrakte Varianzen zusammenfassen.

Die folgende Unterteilung der Verfahren orientiert sich an deren Schichtaufbau. Einschichtsysteme betreiben die Auswertung direkt auf dem Eingabebild. Zweischichtsysteme setzen davor noch eine Vorverarbeitung. Mehrschichtsysteme können auch eine Vorverarbeitung haben, zeigen danach jedoch einen Schichtenstapel, auf dessen Schichten gleiche Auswertungen betrieben werden.

3.1 Einschichtsysteme

3.1.1 Direkte Klassifikation

Eine einfache Methode ein System zur Objekterkennung zu bauen ist, die Pixel direkt mit einem Klassifikator[1] zu verbinden. Jedes Pixel ergibt eine neue Eingabedimension. Schon für kleine Bilder ergeben sich so sehr große Merkmalsräume. Jedes Bild wird als Punkt in dem Merkmalsraum repräsentiert. Die Repräsentation umfasst das gesamte Bild als Kontext. Die interne Repräsentation ist nicht aussagekräftig, da keine höheren/abstrakten Repräsentationen erstellt werden. Allerdings kann jederzeit in das Ursprungsbild zurücktransformiert werden. Es wird auch keine Integration von Informationen vorgenommen.

Der Klassifikator weist jedem gelernten Objekt Bereiche im Merkmalsraum zu. Problematisch ist dabei, dass schon für einfache Varianzen wie Verschiebung und Skalierung der Raum starken Transformationen unterliegt und bezüglich der Varianzen auch kaum generalisiert werden kann.

[1] Z.B. die aktuell erfolgreichen Support-Vector-Machines (Schölkopf and Smola, 2001).

Praktisch müsste jedes Objekt in allen affinen Transformationen und Ausprägungen separat gelernt werden. Dieser Aufwand ist nicht handhabbar.

3.1.2 Globale Auswertung

Bildinformation von Objekten ist in der zweidimensionalen Struktur gegeben. Diese unterliegt Varianzen und ist daher schwer auszuwerten. Ein Ansatz dies zu Umgehen, ist es, Bildeigenschaften global über das Bild auszuwerten. Ein Einschichtsystem kann z.B. direkt Farbhistogramme auswerten. Die Informationsintegration besteht in diesem Fall im Auszählen von Pixeln mit gleichen Farbausprägungen. Kontext wird global ausgewertet und dann als globale Bildstatistik repräsentiert. Informtaionen über die Anordnung des Kontextes gehen verloren und so gibt es für Mustervergleiche anhand der Statistiken großen Spielraum für Mehrdeutigkeiten. Dafür ist das System robust gegenüber affinen Varianzen. Alle höheren Varianzen sind nicht aufzufangen, wenn sie die Farbausprägungen verändern.

Die Metrik von Farbhistogrammen entspricht in der Regel nicht der von Bildinhalten. D.h. Histogramme unterschiedlicher Objektklassen können sich ähnlicher sein, als Histogramme von Bildern der gleichen Klasse.

3.1.3 Mustervergleich

Bei diesen Ansätzen werden Eingabebilder direkt mit einem Musterbild verglichen. Die Differenz zwischen den Bildern kann als Fehlerbild gerechnet werden. Ein Übereinstimmungswert kann invers zu den gemittelten Fehlern gerechnet werden.

Das Muster wird als Bild repräsentiert. Dabei ist der Kontext eindeutig gegeben. Informationsintegration findet statt, indem die Differenzwerte aufsummiert werden. Lernen erfolgt durch Vorgabe von bekannten Mustern.

Das Musterbild kann über ein Eingabbild „geschoben“ werden. Als Ergebnis bekommt man ein neues Bild mit Übereinstimmungswerten. Wenn das Musterbild entsprechend transformiert wird, können auf diese Art und Weise affine Varianzen aufgefangen werden. Dürfen diese in Kombination auftreten, wird der Suchaufwand schnell unhandlich. Ausprägungsvarianzen sind nicht aufzufangen.

3.2 Zweischichtsysteme

Zweischichtsysteme haben zur Zeit die größte Bedeutung in industriellen Anwendungen. Sie bestehen aus einer Vorverarbeitungsschicht und einer

Auswerteschicht. Die Vorverarbeitungsschicht extrahiert meist lokale Strukturbeschreibungen und übergibt sie als „Eigenschaftskarte“ an die Auswerteschicht.

3.2.1 Vorverarbeitungsschicht

Der Auslegung der Vorverarbeitungsschicht kommt eine besondere Bedeutung zu, da sie bestimmt, welche Eigenschaften überhaupt detektierbar sind. In der Regel wird versucht, Bildinformation zu reduzieren - aber nur soweit Muster unterschiedlicher Konzepte noch unterschieden werden können.

Oft wird ein regionaler Mustervergleich auf dem Eingabebild durchgeführt, der „Filter“ genannt wird. Dazu wird eine Faltung mit einer „Filtermaske“durchgeführt, die dem Muster entspricht.

Bei der Größe der Maske ist ein Kompromiss einzugehen (z.B. Mallat, 2009). Sie darf nicht zu groß sein, da die extrahierte Information sonst nur noch spezifisch auf die Maske reagiert und sich die Nachteile von Mustervergleichssystemen einstellen. Wird dagegen die Maske zu klein gewählt, erreicht man kaum eine Informationsreduktion.

Klassische Filter können z.B. zur Linien- und Eckendetektion eingesetzt werden. Sind die Objekte in verschiedenen Skalierungen im Bild gegeben, müssen auch Filter mit unterschiedlichen Maskengrößen eingesetzt werden. Problematisch ist es dann zwischen den Ergebnissen dieser Filter zu generalisieren. Besser funktionieren spezielle multiskalen Detektoren, wie z.B. in Quast and Teichert (2001) für Linien und in Mikolajczyk and Schmid (2004) (Harris-Detektor) für Ecken beschrieben.

Filter können eine regionale Wellenform beschreiben (Gabor- oder allg. Waveletfilter (z.B. Mallat, 2009)). Sie haben eine bestimmte Frequenz und eine Richtung in der Bildebene. Für jede Bildposition wird ein Satz von unterschiedlichen Frequenzen und Richtungen berechnet und als Vektor an die Auswerteschicht weitergereicht. Auch hier ergibt sich das Problem, dass schlecht zwischen verschiedenen Frequenzen und Richtungen generalisiert werden kann.

Aktuell als Vorverarbeitung recht beliebt ist die „Scale Invariant Feature Transform“ (SIFT), die lokale Struktur als Gradienteninformation in einem Vektor bereitstellt. Sie ist positions- und skaleninvariant und durch eine besondere Orientierungszuordnung auch rotationsinvariant (Lowe, 1999). Eine Erweiterung ist in Grenzen auch affin invariant (Lowe, 2004). Die extrahierten SIFT-Vektoren können direkt zu einer Art Datenbankaufruf von gespeicherten (Muster-)Vektoren verwendet werden. Ein Objekt ist detektiert, wenn sich dort auch andere zugehörige Vektoren übereinstimmen.

Meist werden in der Auswertschicht aber komplexere Detektionsmethoden verwendet. Darauf wird noch in Abschnitt 3.2.6 eingegangen.

Ähnlich den SIFT-Detektoren lassen sich Eigenschaftdetektoren einsetzen, die Grenzen eines lokalen Bereichs so einstellen, dass die darin enthaltene Bildinformation maximiert wird. Auch sie liefern positions-, skalen-, und rotationsinvariante Beschreibungsvektoren (Kadir and Brady, 2001), die in einer Erweiterung auch affin invariant sind (Kadir et al., 2004).

Als Vorverarbeitung können auch lokal berechnete statistische Momente verwendet werden (z.B. Alferez and Wang, 1999; Rodrigues, 2000; Amit, 2002). Dalal and Triggs (2005) setzen auch lokale Histogramme von Richtungsinformation ein.

3.2.2 Direkte Klassifikation und Mustervergleich

Auf die extrahierten Bildeigenschaften kann wieder ein einfacher Klassifikator gesetzt werden, der die Ergebnisse der Vorverarbeitung direkt auswertet. Sind diese in einem Bildraster gegeben, können auch wieder Maskenvergleiche darauf durchgeführt werden. In dieser Art sind viele aktuelle Bilderkennungssysteme in der Industrie aufgebaut (z.B. Bernd Jähne, 2000; Sonka and Hlavac, 2007; Stemmer, 2011).

Bedingt durch die Vorverarbeitung werden meist regionale Bildbereiche (Teilmuster) repräsentiert und verglichen. Die Informationsintegration erfolgt durch die Vorverarbeitungsschicht. Der Kontext bleibt regional starr. Das ist z.B. zum Überprüfen von Druckmustern in Produktionsprozessen nicht hinderlich.

Lernen erfolgt durch Speichern von Vorverarbeitungergebnissen von vorgegebenen Mustern.

Je nach Vorverarbeitung können klassische Varianzen in geringen Grenzen aufgefangen werden. Ausprägungsvarianzen sind so nicht realisierbar.

Alle Verfahren, die einen Merkmalsraum in Bereiche aufteilen, haben das Problem, dass transformierte Teilbilder nicht wiederverwendet werden können. Für das Erkennen von Hausfassaden, wäre es z.B. hilfreich, wenn man Detektoren für einzelne Fenster hätte, die an unterschiedlichen Positionen nutzbar wären.

3.2.3 Globale Auswertung

Viele aktuelle Verfahren zur Bildähnlichkeitssuche in großen Bildbeständen verwenden globale Eigenschaften der extrahierten Vorverarbeitungsergebnisse (Lew, 2006; Datta et al., 2008).

Einige Verfahren gewichten die Ergebnisse von unterschiedlichen Vorverarbeitungskanälen (Jacobs et al., 1991; Mel, 1997), um möglichst spezifische Objekterkenner zu realisieren. Andere Verfahren nutzen quasi-regionale Eigenschaften, indem sie das Bild in ein festes Gitter (z.b. oben, mitte, unten, rechts und links) unterteilen. Innerhalb der Gitterfelder wird wieder eine globale Analyse vorgenommen (Lew, 2006).

Moderne Verfahren verwenden als Vorverarbeitung affin invariante Eigenschaftsdetektoren (Csurka et al., 2004; Fei-Fei and Perona, 2005; Zhang et al., 2007) und erhalten sich damit eine gewisse regionale Strukturspezifität.

Bei den globalen Zweischichtverfahren bestehen Muster- und Eingabebildrepräsentationen aus Eigenschaftsvektoren oder Histogrammen der Vorverarbeitungsergebnisse. Für Strukturen und Anordnungen kann es regionale Beschreibungen geben, die jedoch global nicht mehr zugeordnet werden können und in dem Sinne auch wenig aussagekräftig sind.

Die Informationsintegration erfolgt durch Bilden der Eigenschaftsvektoren oder Histogramme. Dabei wird Kontext global ausgewertet. Information über die Anordnung der Vorverarbeitungsergebnisse geht dabei verloren und es entstehen die schon in Abschnitt 3.1.2 beschriebenen Mehrdeutigkeiten. Bekannte Muster werden auch als Eigenschaftvektor oder Histogramm repräsentiert. Bei der Detektion kann ein einfacher Vergleich, z.B. eine Korrelationsmessung verwendet werden.

Es liegt in der Natur der globalen Auswertung, dass affine und topologische Varianzen aufgefangen werden können. Allerdings leidet darunter die Detektionspräzision. Helligkeits- und Farbvarianzen können in Abhängigkeit der Vorverarbeitung zu Problemen führen. Ebenso führen abstrakte Varianzen zu Unterschieden in den Bildinhalten, die sich im Allgemeinem auch auf die Eigenschaftsvektoren oder Histogramme auswirken. Daher können diese Verfahren solche Varianzen auch nur begrenzt auffangen.

3.2.4 Symbolische Auswertung

Symbolische Auswertung zeichnet sich durch die Verwendung von Konzepten und deren Relationen aus. So kann Bildinformation von Mustern in Graphen repräsentiert werden, die Bildprimitive (Ergebnisse aus der Vorverarbeitung) und deren geometrische Relationen beschreiben (Fischler and Elschlager, 1973; Niemann, 1990; Felzenszwalb and Huttenlocher, 2005). Die Verfahren stellen einen Bezug zwischen Sensorinformation und symbolischer Verarbeitung her („Symbol Grounding“ Harnad, 1990). Daher sind innere Repräsentationen Konzepten zuzuordnen und damit sehr aussagekräftig.

Ist symbolische Information über Inhalte und Anordnungen extrahiert worden, können Methoden der klassischen Künstlichen Intelligenz eingesetzt werden, um daraus neues Wissen abzuleiten. Die Informationsintegration kann z.B. anhand von Produktionsregeln erfolgen. Kontext kann dabei ausgewertet und für folgende Auswertungen erhalten bleiben. Objekteigenschaften können direkt als Attribute an die Objektrepräsentation gebunden werden. Die Kombinierbarkeit kann durch erlerntes Modellwissen beschränkt sein.

Das Einstellen des Modells kann explizit per Hand oder auch durch Beispielbilder mit ausgezeichneten (Teil-)Objektregionen erlernt werden.

Es ist relativ leicht, das Modell so einzustellen, dass eine klassische Varianz aufgefangen werden kann. Für das Auffangen mehrerer Varizen wird der Aufwand jedoch schnell sehr groß. Abstrakte Varianzen sind nur durch explizites Vorgeben entsprechender Modelle möglich. Generalisieren zwischen unterschiedlichen Modellen fällt im Allgemeinen schwer. Problematisch ist die stabile Zuordnung von Symbolen zu Bildstrukturen unter Varianz. Es ist fraglich, ob das je mit symbolischen Methoden zu lösen ist, oder ob das nur mit subsymbolischen Zwischenrepräsentationen gelingen kann (Feldman and Ballard, 1982).

3.2.5 Dynamic Link Matching

Das Verfahren „Dynamic Link Matching“ (DML) wurde entwickelt, um Objekte mit Ausprägungsvarianzen zu Erkennen (Lades et al., 1993; Konen et al., 1994). Die Topologie der Objektansichten muss aber relativ konstant bleiben und daher wurde das Verfahren hauptsächlich zur Gesichtserkennung (Würtz, 1995) und mit Erweiterungen auch zur Gesichtidentifikation eingesetzt (Wiskott et al., 1997).

An bestimmten Bildpunkten werden regionale Eigenschaften durch eine Vorverarbeitung mit einem Satz von Gaborfiltern extrahiert. Diese werden mit Modellrepräsentationen verglichen, die durch Musterbilder gelernt wurden. Die Modellrepräsentationen besitzen untereinander Verbindungen, die relative Anordnung beschreiben. Sie werden im Erkennungsvorgang durch ein Relaxationsverfahren mit den Anordnungen im Eingabebild in Übereinstimmung gebracht. So entsteht iterativ eine Korrespondenzkarte.

Es werden lokale Frequenzen und deren Orientierungen auf verschiedenen Skalen als Gaborjets repräsentiert. Diese werden aber nur an Punkten eines relativ groben Rasters und im Fall der Gesichtserkennung an markanten Punkten im Gesicht gespeichert. Diese Stellen lassen sich in einem Eingabebild benennen und dort lässt sich auch die typische Bildstruktur

aus den Filterwerten wieder rekonstruieren. So ist die innere Repräsentation relativ aussagekräftig.

Informationsintegration wird nur im Rahmen der Vorverarbeitung durchgeführt. Die Kontextauswertung umfasst die präzise Repräsentation im Bereich der Gabor-Filtermasken und die relativ flexiblen überregionalen Repräsentationen der Jet-Anordnungen. Modell- und Eingabebildrepräsentationen werden dynamisch aneinander gebunden. Eigenschaften, die nicht als Jet-Repräsentation im Modell enthalten sind, können aber nicht gebunden werden. So finden in einem Gesicht Augen und Mund in Modell und Eingabebild zueinander - die Ausprägung einer Kopfbedeckung kann aber nicht repräsentiert werden.

Lernen erfolgt durch Mustervorgaben und eventuell noch Vorgaben der Jet-Positionen.

Durch die flexible Struktur können klassische Varianzen recht gut aufgefangen werden. Die Topologie darf aber nur gering verzerrt sein. Das Verfahren kann keine Teilobjekte an unterschiedlichen Positionen erkennen. Abstrakte Varianzen sind nicht auffangbar.

3.2.6 Auswertung affin invarianter Vorverarbeitung

Für viele aktuelle Auswertverfahren werden affin invariante Detektoren eingesetzt[2]. Zunächst wurden sie zur Bestimmung von Punktkorrespondenzen genutzt (Klette et al., 1998; Hartley and Zisserman, 2004). Dann konnten Bildregionen zugeordnet werden (Ferrari et al., 2004), wobei auch die Objektlage bestimmt werden konnte (Gordon and Lowe, 2006).

Mit diesen Detektoren können bekannte Objekte (und nur diese) mit affinen Varianzen in Szenen wiedererkannt werden. Höhere Varianzen können zum Teil mit global auswertenden Methoden aufgefangen werden (Csurka et al., 2004; Fei-Fei and Perona, 2005; Zhang et al., 2007). Deren Eigenschaften und Nachteile wurden schon in Abschnitt 3.2.3 beschrieben. Der fehlende Ortsbezug wurde dann mit zusätzlichen wahrscheinlichkeitsbasierten Verfahren (Fei-Fei et al., 2004; Sudderth et al., 2005; Schmid, 2006; Niebles and Fei-Fei, 2007) oder hierarchischen Verfahren (Schmid, 2006; Grauman and Darell, 2007b) zum Teil wiederhergestellt.

Dann wurden Objekte explizit in ein Wahrscheinlichkeitsmodell übernommen (Carneiro and Lowe, 2006) und zusätzlich auch deren Ausprägung (Fergus et al., 2003, 2007; Felzenszwalb et al., 2008). Es wurden auch Modelle für eine hierachische Objektrepräsentation vorgestellt (Bouchard and

[2]Haupsächlich von Lowe (1999, 2004) und Kadir and Brady (2001); Kadir et al. (2004) siehe Abschnitt 3.2.1

Triggs, 2005; Fidler et al., 2007; Sudderth et al., 2008) - auch für konturbasierte Repräsentation (Opelt and Zisserman, 2006; Ravishankar et al., 2008).

Für Objekte mit Ausprägungsvarianzen von Objektteilen, kann in Musterbildern der zugehörigen Klasse einzeln nach den Objektteilen gesucht werden (Boiman et al., 2008).

Die folgende Bewertung soll für die modernen Verfahren erfolgen, die Objektteile erkennen können, da diese Verfahren am besten geeignet sind, Ausprägungsvarianzen aufzufangen.

Es werden Wahrscheinlichkeitsfunktionen (meist nach Bayes) bestimmter Anordnungen, Ausprägungen und (Teil-)zugehörigkeiten repräsentiert. Für die Modelle werden oft Gaußverteilungen genommen, die einfach zu parametriesieren sind. Innere Repräsentationen sind nur bedingt aussagefähig, da nur bei einigen Verfahren die (Teil-)Repräsentationen in Bildausschnitte zurückgewandelt werden können (z.B. Fergus et al., 2003).

Die Vorverarbeitung liefert meist Eigenschaftsvektoren, die aus Histogrammen lokaler Gradienten (HOG) gewonnen werden. Diese Werte werden mit dem Wahrscheinlichkeitsmodell verglichen. Eine Informationsintegration erfolgt nur in der Vorverarbeitung. Kontext wird regional per Vorverarbeitung ausgewertet. Überregionale Kontextauswertung erfolgt anhand des Wahrscheinlichkeitsmodells. Ausprägungen von (Teil-)Objekteigenschaften werden nicht gebunden, können aber recht flexibel parametrisiert werden.

Im Erkennungsvorgang wird das Wahrscheinlichkeitsmodell zum Vergleich verwendet, das zuvor aus Vorverarbeitungsergebnissen von trainierten Bildern erstellt wurde. Das Lernen von Objektzugehörigkeiten und Ausprägungsvariazen gelingt aus kombinatorischen Gründen im Moment nur für wenig Teilobjekte.

Da die Vorverarbeitung bereits affin invariant ist, sind es auch die Ergebnisse. Ausprägungsvarainzen können in gewissen Grenzen aufgefangen werden. Dabei generalisieren die Verfahren recht gut.

Dies ist möglich, solange eine Anwendung stets ein Objektkonzept aus einer Konzeptmenge auswählen muss. Schwierig wird es, wenn bisher nicht bekannte Teilobjektkombinationen zu erkennen sind oder auch Szenen, die kein gelerntes Objekt enthalten (Szeliski, 2010).

3.3 Mehrschichtsysteme

Mehrschichtsysteme sind in dem Sinne biologisch motiviert, als der Schichtenstapel den biologischen Verarbeitungspfad nachstellen soll und dabei eine Komplexitätsreduktion der Bildinformation vorgenommen wird (Logothe-

tis, 1998). Eine schichtenbasierte Verarbeitung bietet auch die Möglichkeit, Repräsentationen auf unterschiedlichen Komplexitätsstufen flexibel für variierende Kombinationen zu halten (Oram and Perrett, 1994).

3.3.1 Shifter Circuit

Die Idee, mit sogenannten „Shifter-Circuits“ Bildinhalte zuerst in eine Normalform zurück zu transformieren und dann zu erkennen, stammt von Anderson and van Essen (1987) und wurde von Olshausen et al. (1993, 1995) sowie Postma et al. (1997) umgesetzt. Es gibt auch Einschichtsysteme, die Rücktransformation mithilfe von Gestaltprinzipien realisieren (Shen and Horace, 1997).

Bildinformationen werden über mehrere Schichten weitergereicht. Die Verbindungen können dabei gesteuert werden. So können Bildinhalte verschoben und auch skaliert werden. Auf der höchsten Schicht soll eine normierte Ansicht der Bildinformation zu finden sein, die dann mit klassischen Techniken klassifiziert werden kann. Die Steuerung der Verbindungen übernimmt ein Kontrollblock, der die Ansteuerung der Verbindungen aus den Bildinformationen ableitet.

Auf den Schichten werden Zwischenstadien des Routingprozesses repräsentiert. Bildinhalt wird nur durchgeleitet. Es gibt keine szenenbeschreibenden Repräsentationen, die aus dem Eingabebild abgeleitet werden. Innere Repräsentationen sind aussagekräftig in dem Sinne, dass stets ein (Teil-) transformiertes Bild zur Verfügung steht. Eine Segmentierung und Dekomposition ist aber nicht durchführbar.

Informationsintegration findet nicht im o.g. Sinn statt. Wenn die Verbindungen zwischen den Schichten so gestellt sind, dass Bildregionen runterskaliert werden, findet ledigleich eine Mittelung statt. Eine Kontextauswertung wird nur vom Kontrollblock durchgeführt - das auch nur auf stark tiefpassgefilterten Bildinformationen. Das Einstellen der Verbindungen kann auch als dynamisches Binden von Positionsinformation an Bildinformation gedeutet werden. Bekanntes Wissen ist in der Endklassifikation und in der Verbindungssteuerung gegeben.

Dieses wird mit einem überwachten Lernverfahren für bestimmte Eingabebildstrukturen im Kontrollblock erlernt. Problematisch ist das Behandeln nicht bekannter Muster. Für den Kontrollblock wird die Bildinformation zwar mit einer groben Unschärfe belegt. So sind Ziffern immer noch gut zu erkennen - beliebige Muster müssen jedoch immer wieder neu gelernt werden. Das Erlernen der Endklassifikation kann mit klassischen Techniken erfolgen.

Das Shifter-Circuit-Verfahren kann Verschiebungs- und Skalierungsvarianzen auffangen. Theoretisch sind auch andere klassischen Varianzen denkbar. Ein Problem stellt dann aber die Rücktransformation von Bildregionen dar, deren Inhalt in Kombination mit möglichen Transformationen immer mehrdeutiger wird. Die eigentliche Objekterkennung wird auf die oberste Schicht verschoben. Damit wird das Erkennen von abstrakten Varianzen auch nur auf diese Schicht verschoben und entspricht dann dem, was für Einschichtsysteme schon in Abschnitt 3.1.1 beschrieben wurde.

3.3.2 Neocognitron

Basierend auf dem Cognitron (Fukushima, 1975) stellte Fukushima (1980) das Neocognitron vor. Es ist ein Mehrschichtverfahren, das abwechselnd strukturspezifische Schichten und Schichten zur regionalen Summation (Pooling) hintereinander schaltet. Durch die Summation wird in Richtung höherer Schichten die Positionsinformation immer gröber und kann mit abnehmenden Auflösungen repräsentiert werden. Im Gegenzug werden mehr (Teil-)Objekt-spezifische Schichten eingeführt. Am Ende der Verarbeitung steht für jedes zu unterscheidende Konzept eine dedizierte Schicht.

Das Verfahren wurde in spezialisierten Formen erfolgreich zur Erkennung von handgeschriebenen Ziffern eingesetzt (Fukushima et al., 1983; Fukushima, 1988; LeCun et al., 1989, 1990).

Repräsentiert werden Objekte, indem auf jeder Schicht die zugehörigen (Teil-)Spezifitäten aktiviert sind. Jede Spezifität beschreibt auch Anordnungen von Bildstrukturen, die auf unteren Schichten noch direkt den Strukturen des Eingabebildes zugeordnet werden können. Für Folgeschichten ist eine Anordnung die zu einer Aktivität führte nicht mehr sichtbar, was insbesondere für Rotationen problematisch ist. Die Teilrepräsentationen sind nicht aussagekräftig in dem Sinne, dass man dort Teilobjekte sehen könnte. Man kann die räumlich zuzuordnenden Aktivierungen aber gut im Schichtenstapel zurückverfolgen und nachvollziehen wie eine Detektion zustande gekommen ist (LeCun et al., 2004). So ist im Prinzip auch eine Dekomposition von Objektteilen möglich. In der Praxis lassen sich aber nur relativ wenig Schichten realisieren. Daher müssen die Masken der Spezifitäten relativ groß sein, um für eine Objektdetektion ausreichend Bildbereich abzudecken. Zur detaillierten Segmentierung von Teilobjekten bräuchte man kleine Masken und dann viele Schichten.

Informationsintegration wird in den Neocognitronverfahren durch regionale Kontextauswertung mit Masken vorgenommen. Objektattribute werden in diesem Verfahren nicht an die Repräsenationen gebunden. Durch die regionale Summation nach dem Detektieren, ist jedoch eine gewisse Frei-

heit für unterschiedliche Kombinationen von Teilrepräsentationen gegeben. Riesenhuber and Poggio (1999a) ersetzen die Summation durch eine Maximumfunktion und sehen darin eine Lösung für flexibles Binden von Teilrepräsentationen (Riesenhuber and Poggio, 1999b). Die Funktion ist biologisch motiviert und soll das Verhalten von „Komplexen Zellen“ im visuellen Kortex nachstellen (Tsunoda et al., 2001; Gawne and Martin, 2002). Das Neocognitronverfahren arbeitet rein vorwärtsgerichtet. Bekanntes Wissen ist in den Masken der Spezifitätsschichten gegeben.

In der klassischen Form des Neocognitrons wird ein unüberwachtes Lernverfahren verwendet, welches für die Spezifitäten eine „Winner-Take-All“ - Bekräftigung verwendet (Fukushima, 1980). Für die praktischen Anwendungen zur Handschriftenerkennung mussten jedoch die Schichten einzeln mit einem überwachten Lernverfahren eingestellt werden (Fukushima, 1988). Gelernte Schichten bleiben dann für das Erlernen der Folgeschicht unverändert. Eine analytische Betrachtung der Lernverfahren und verschiedene Leitstungsmessungen zeigt Lovell et al. (1997).

Verschiebungsinvarianz ergibt sich für das Neocognitron durch gemeinsam verwendete Maskeneinstellungen der Spezifitätsschichten. Eine relativ geringe Skalierungsinvarianz ergibt sich aus der regionalen Summation. Diese verbessert sich etwas, wenn man statt der Summenbildung eine Maximumfunktion verwendet (Riesenhuber and Poggio, 1999a; Serre et al., 2002, 2005). Rotationsinvarianz ist für das Neocognitron nur in sehr kleinem Umfang gegeben. Es gibt Verfahrenserweiterungen, die das Erkennen von rotierten Bildobjekten ermöglichen. Dabei potenziert sich aber der Suchaufwand (Satoh et al., 1997) oder es wird eine separates Netzwerk benötigt, um die Rotation zu erkennen, wobei das die erkennbaren Muster einschränkt (Fukumi et al., 1997; Satoh et al., 1999). Topologische Invarianz ist in Grenzen möglich. Das Verfahren lernt dann mehrere Ansichten eines Konzeptes. Wenn sich diese stark unterscheiden, wird sich auf hohen Ebenen eine Art „Oder-Funktion“ ergeben. Weitere Ansichten werden dann eher nicht gut generalisieren. Bessere Eigenschaften zeigen da Verfahren, die wie das Neocognitron Maskenspezifitäten und regionale Summation verwenden - die Spezifitäten aber frei in der Schichtenhierarchie einsetzen können (Teichert and Malaka, 2002, 2003). Abstrakte Varianzen zeigen beim Neocognitron ähnliche Probleme wie topologische. Generell verbessert man die Fähigkeiten, solche Varianzen aufzufangen, indem man auf hohen Schichten viele (Zwischen-)Konzepte zulässt (LeCun et al., 2004).

3.3.3 VisNet

Das bilogisch motiviertes Mehrschichtsystem wurde von Wallis and Rolls (1997) und Elliffe et al. (2002) vorgestellt. Gegenüber den bisher vorgestellten Mehrschichtverfahren benutzt es keine gemeinsam genutzten Verbindungseinstellungen, um Verschiebungsinvarianz zu erreichen. Um noch besser Verschiebungs- und Skalierungsvarianzen auffangen zu können, wurde später noch ein extra Verarbeitungspfad eingeführt, der die Verbindungen beeinflusst und so ein Aufmerksamkeitssystem darstellen soll (Deco and Rolls, 2004; Rolls and Stringer, 2006).

Wie beim Neocognitron, sind Repräsentationen auch über Aktivierungspfade gegeben, die mehrere Schichten umfassen. VisNet ist im Prinzip besser in der Lage, Information über Anordnungen weiter an höhere Schichten weiterzuleiten, wenn dies zur Unterscheidung von verschiedenen Objekten notwendig ist. Das liegt an dem unüberwachten Lernverfahren der kompetetiven Schichten, das wesentliche Merkmale bekräftigt. Wie beim Neocognitron sind Dekompositionen im Prinzip möglich.

Die Informationsintegration funktioniert ausschließlich über Kontextauswertung anhand der Verbindungen zwischen den Schichten. Es gibt keine regionale Summation. Das Verfahren arbeitet vorwärtsgerichtet.

Zur Erlangung von Verschiebungsinvarianz wird das Eingabebild an verschiedenen Positionen gezeigt. Die Aktivierung der obersten Schicht wird beibehalten. Dann werden alle Verbindungen, die sich zwischen aktivierten Schichtelementen befinden, verstärkt (Hebbsches Lernen).

Im Prinzip lassen sich so beliebige Invarianzen erlernen. Auch Ausprägungsvarianzen sind möglich. Allerdings wird der Lernaufwand dafür durch die kombinatorische Vielfalt schnell zu groß.

VisNet ist aus verfahrensarchitektonischer Sicht komplementär zum Shifter-Circuit. Das Shifter-Circuit führt eine Objektabbildung auf eine einzelne normalisierte Repräsentation zurück, während VisNet eine Repräsentation erstellt, die alle Varianzen in sich trägt. Weder VisNet noch Shifter-Circuit konnten bisher auf größere Bilderserien angewendt werden. Sie markieren aus verfahrensarchitektonischer Sicht zwei Extrema, die wahrscheinlich keine große Anwendungsrelevanz mehr erlangen. Erfolgreiche Verfahren wie die um das Neocognitron beschreiten einen Mittelweg. Sie bilden einen Verarbeitungspfad, auf dem ein Teil der Information über Varianzen reduziert wird und ein anderer Teil so transformiert wird, dass er einer Standardrepräsentation entspricht.

3.4 Diskussion und Zusammenfassung

Aktuellen Verfahren zur Bilderkennung erreichen noch keine hohe Konzeptabstraktion. Objektabbildungen mit abstrakten Varianzen sind daher noch nicht richtig zu erkennen.

Einigen Verfahren gelingt dies nur für trainierte Bilder. Wenn dann Testbilder mit abstrakten Varianzen dazukommen, können die Verfahren nicht auf dem gewünschten Abstraktionsniveau generalisieren.

Am erfolgreichsten sind die Verfahren mit affin invarianter Vorverarbeitung. Ihre Wahrscheinlichkeitsmodelle bieten jedoch nur begrenzt Möglichkeit auch Repräsenationshierarchien zu realisieren.

Daher ist es wichtig, Verarbeitungsmethoden zu entwickeln, die grundsätzlich neue Methoden verwenden, um die gewünschten Eigenschaften zu erreichen.

Kapitel 4

Provadero-Verfahren

In diesem Kapitel soll das Provadero-Verfahren vorgestellt werden. Es stellt eine konzeptionelle Lösung für die Realisierungsfrage in der wissenschaftlichen Fragestellung dar. Im folgenden Kapitel 5 wird eine konkrete Realisierung des Verfahrens vorgestellt.

Mit der Entwicklung des Provadero-Verfahrens soll der Versuch unternommen werden, wesentliche Funktionsprinzipien der biologischen Bildanalyse zu modellieren. Wie schon in Kapitel 2.2 dargestellt wurde, ist der bisher zusammen getragene Satz von Funktionsprinzipien unvollständig und nicht validiert. Daher ist die Vorgehensweise in diesem Beitrag konstruktiv, d.h. es werden Funktionsweisen postuliert, realisiert und dann im übernächsten Kapitel auf Relevanz überprüft. Es werden verschiedene Module definiert, die die Funktionen ermöglichen sollen. Diese Module werden in Lern- und Assoziationsprozessen iterativ eingesetzt.

In klassischen Verarbeitungsmodellen werden typischerweise Aktivitäten als skalare Werte modelliert. Für das hier vorgestellte Verfahren wird zusätzlich eine skalare Größe verwendet, die eine Ausprägung beschreibt. So sind die verwendeten Module leider auch nicht einzeln in klassischen Umgebungen zu evaluieren. Nur einige Module sind separat zu evaluieren. Der Rest muss als Gesamtheit ausgewertet werden. Das macht es schwierig, Realisierungen zu optimieren. Es wurde daher Wert darauf gelegt, die Module möglichst einfach zu gestalten und dabei auch nur wenige Parameter einzuführen. Entwicklung und Evaluierung des Verfahrens entfernen sich dennoch vom klassischen Vorgehen, schrittweise nur kleine Änderungen einzuführen und diese separat zu evaluieren. Die Module wurden im Entwurfs- und Evaluierungsprozess oft verändert. In den folgenden Beschreibungen wird nur der letzte Entwicklungsstand dargestellt. Teilweise werden verworfene Realisierungen jedoch noch als Varianten diskutiert.

4.1 Repräsentation

4.1.1 Was wird repräsentiert?

Mit dem Provadero-Verfahren soll der Versuch unternommen werden, Varianzen explizit zu repräsentieren. Die Detektion soll nicht nur Varianzen auffangen, sondern auch einen Wert über deren Ausprägung liefern. So sollen Folgedetektionen spezifisch auf bestimmte Ausprägungen reagieren können. Das ist wichtig, wenn ein Teilkonzept unterschiedlich für Folgedetektionen benutzt werden soll. Soll z.B. das Teilkonzept „Kreisrunde Struktur“ für die Erkennung von Autorädern und Verkehrszeichen eingesetzt werden, spielt die Information über die Farbe auf dem Rand eine wichtige Rolle. Klassische Detektionselemente berechnen nur einen Wert, der die Übereinstimmung mit dem Teilkonzept beschreibt. Im beschriebenen Beispiel würde also das Teilkonzept „Runde Struktur“ detektiert. Information über die Ausprägung der Farbe stünde für Folgedetektionen nicht mehr zur Verfügung.

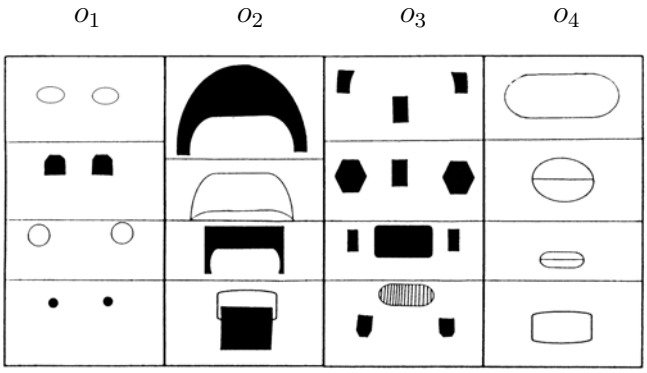
Ein klassisches Detektionselement reagiert spezifisch auf eine bestimmte (Teil-)Bildstruktur. Es besitzt typischerweise einen skalaren Signalausgang y, der in diesem Fall eine Aktivierung größer Null zeigt. Die Information über die Ausprägung der Bildstruktur wird verworfen.

Ein Provadero Element hat dagegen zwei Signalausgänge (Abbildung 4.1). Ein Signal ist die (skalare) Konfidenz c, die ebenfalls auf eine (Teil-)-Bildstruktur spezifisch mit einem Ausgangswert größer Null reagieren soll. Zusätzlich wird aber auch noch eine Ausprägung v berechnet, die die Ausprägung der Bildstruktur beschreiben soll. Die (skalare) Ausprägung kann Werte größer, kleiner und gleich Null annehmen. Welche Ausprägung damit konkret repräsentiert wird, hängt davon ab, welche Muster zuvor trainiert wurden. Der Nullwert repräsentiert ein gemitteltes Trainingsmuster. Welche (Teil-)Bildstrukturen repräsentiert werden, wenn der Wert ins Positive oder Negative geht, hängt davon ab, welche Bildstrukturen der Trainingsmuster die größte Varianz gezeigt haben.

Elemente, die Varianzen explizit am Ausgang repräsentieren, wurden schon von Poggio and Edelman (1990) und Liu et al. (1995) eingesetzt. Allerdings waren das Zweischichtsysteme, die nur einfache klassische Transformationen berechnen konnten.

4.1.2 Wie werden Strukturen und Anordnungen kodiert?

Viele Ansätze versuchen Eigenschaften der frühen rezeptiven Felder in biologischen Sehsystemen nachzuahmen. Dabei bleibt meist offen, wie höhere



Klassische Detektion:				
Aktivierung:	$y_1 > 0$	$y_2 > 0$	$y_3 > 0$	$y_4 > 0$
Provadero Detektion:				
Konfidenz:	$c_1 > 0$	$c_2 > 0$	$c_3 > 0$	$c_4 > 0$
Provadero Detektion:				
Ausprägung:	$v_1 > 0$	$v_2 > 0$	$v_3 > 0$	$v_4 > 0$
	$\uparrow$	$\uparrow$	$\uparrow$	$\uparrow$
	$v_1 = 0$	$v_2 = 0$	$v_3 = 0$	$v_4 = 0$
	$\downarrow$	$\downarrow$	$\downarrow$	$\downarrow$
	$v_1 < 0$	$v_2 < 0$	$v_3 < 0$	$v_4 < 0$

Abbildung 4.1: Detektion mit klassischem- und Provadero-Element. Es sind die Muster aus Abbildung 2.1 gezeigt. Jede Spalte steht für ein unterschiedliches Konzept o. In den Zeilen findet man unterschiedliche Ausprägungen der Bildstruktur für diese Konzepte. Eine klassische Detektion ist auf ein bestimmtes Konzept spezifisch. Liegt eine der gezeigten Abbildungen des Konzeptes als Eingabesignal an, so liefert der Ausgang y des dafür spezifischen Detektors typischerweise eine Aktivierung größer Null. Die Provadero Detektion besitzt zwei skalare Ausgangsgrößen. Die Konfidenz c ist konzeptspezifisch - wie bei der klassischen Detektion. Die Ausprägung v liefert einen Wert, der die Ausprägung der Bildstruktur beschreibt. Sie soll unterschiedlichen Ausprägungen unterschiedliche Werte zuweisen. Die Zuordnungsfunktion orientiert sich an der größten Varianz, die in den Musterbildern gegeben ist.

Verarbeitungszentren diese Ergebnisse weiterverarbeiten. Hier soll versucht werden, eine einheitliche Verarbeitungsfunktion zu realisieren, die entlang des visuellen Verarbeitungspfades gleichermaßen eingesetzt werden kann.

Dies wird im Provaderoverfahren durch Verarbeitungsschichten nachgeahmt. Sie bilden den Verarbeitungspfad durch einen Schichtenstapel nach. Die unterste Schicht im Stapel repräsentiert das Eingabebild. Die höheren Schichten des Stapels repräsentieren zunehmend abstraktere konzeptuelle Bildinhalte. Auf jeder Schicht l liegen Raster mit Rasterpunkten, die durch Positionen p eindeutig zu lokalisieren sind. Die Raster sind auf allen Schichten gleich, d.h. die Rasterauflösung entspricht auf allen Schichten denen des Eingabebildes. So muss keine Auflösungsreduktion durchgeführt und parametrisiert werden.

Jede Schicht ist einem Konzept o zugeordnet. Repräsentationen auf einer Schicht beschreiben die Anordnung dieses Konzeptes in einer regionalen Umgebung. Diese Repräsentationen lassen sich für alle Schichten in einem Zustandsvektor $\mathbf{x}(p)$ an einer Position p zusammenfassen. In diesem Vektor sind also Beschreibungen über die regionalen Anordnungen aller Konzepte gegeben. Auch wenn die Beschreibungen für die einzelnen Konzepte in der hier gewählten Realisierung relativ einfach sind, können durch die Kombination von vielen Konzepten komplexe Bildinhalte beschrieben werden.

An jedem Rasterpunkt wird der Zustandsvektor durch ein Detektionselement ausgewertet. Die Auswertung erfolgt anhand der Spezifität des Detektionselementes. Sie ist innerhalb einer Schicht einheitlich für alle Detektionselemente $\mathbf{s}(l)$ und sie ist charakteristisch für das der Schicht zugeordnete Konzept o.

4.1.3 Sind innere Repräsentationen aussagekräftig?

Die konstanten Rasterauflösungen ermöglichen es, Repräsentationen direkt Bildinhalten zuzuordnen. So kann man für jedes Konzept o nachvollziehen, wie stark es an einer Rasterposition p aktiviert ist $(c(l(o), p))$ und welche Ausprägung es dort hat $(v(l(o), p))$. Es ist auch nachvollziehbar, welche Komponenten von unterliegenden Schichten zur Erstellung einer Repräsentation beigetragen haben. So sollen komplexe Konzepte durch rückwärtgerichtetes Überprüfen segmentierbar sein.

4.2 Integration von Repräsentation

Zum besseren Verständnis der Provadero-Funktionen wird von der in Abschnitt 2.1 entwickelten Gliederung abgewichen und die Beschreibung der

Kontextauswertung vorgezogen.

4.2.1 Wie wird Kontext ausgewertet?

In Abschnitt 3.2.1 wurde schon auf typische Vorverarbeitungsschichten eingegangen. Sie haben den prinzipiellen Nachteil, dass sie separat vom Klassifikationsverfahren entworfen werden müssen. Es gibt im Allgemeinen keine analytische Handhabe für den Entwurf einer guten Kombination. Welche Informationen die Vorverarbeitung detektiert, welche sie verwirft und welche Auflösungsreduktion erfolgt, ergibt sich in der Praxis meist aus Heuristiken. Diese Verfahrensvorgaben werden meist nicht mehr durch einen musterspezifischen Lernvorgang beeinflusst und sind daher im Allgemeinen suboptimal.

Das Provadero-Verfahren umgeht diese Schwierigkeiten und extrahiert direkt auf den Rasterdaten der Repräsentationsschichten, wobei die unterste Rasterschicht das Eingabebild repräsentiert.

Gegenüber klassischen Verfahren, die Kontextinformationen ausgehend von einem Rasterpunkt analysisieren, wird im Provadero-Verfahren der umgekehrte Weg gegangen. Mit einem Diffusionsverfahren $\mathcal{H}$ wird Information über die Anordnung von Ausprägungen über eine Schicht l verteilt und ist dann an jedem Rasterpunkt p abgreifbar. Diese Kontextauswertung wird auf jeder Schicht gleich ausgeführt.

Das Diffusionsverfahren erstellt Strukturbeschreibungen. Das sind Informationen über die Anordnungen der von den Detektionselementen berechneten Ausprägungen $v(l,p)$. Die Strukturbeschreibungen umfassen den dominanten Gradienten $\mathbf{g}(l,p)$ der Ausprägungen in der Umgebung von p sowie dessen Entfernung (Translation) $\mathbf{r}(l,p)$ von der Postion p. Zusätzlich sind beiden Vektoren noch die Konfidenzwerte $c^g(l,p)$ und $c^r(l,p)$ zugeordnet. Diese beschreiben, wie verlässlich die Vektoren berechnet werden konnten. Abbildung 4.2 zeigt eine mögliche Anordnung dieser Werte.

Im Abschnitt 4.1.2 wurde schon angedeutet, dass sich im Provadero-Verfahren Informationen über Anordnungen nicht nur aus den Strukturbeschreibungen einer Schicht und damit für ein (Teil-)Konzept ergeben. Sie werden über viele Schichten zu einem Zustandsvektor zusammengetragen. Welche Schichten dazu einen Beitrag liefern, bestimmt eine Freigabefunktion $\mathcal{E}$. Sie wählt aus, welche Signale von welchen Schichten von den Provadero Detektionselementen ausgewertet werden. Die Freigabe ist ein wichtiger Bestandteil des Provadero-Verfahrens. Bei klassischen Verfahren ist in der Regel vorab festgelegt, wie der Merkmalsraum beschaffen ist, der zur Vorverarbeitung und zur Detektion verwendet wird. Typischerweise hat auch

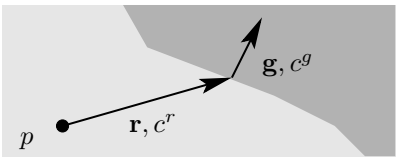


Abbildung 4.2: Repräsentation von Bildstruktur durch zwei Vektoren. Für die Provadero-Architektur wurde ein einfaches Kontextmodell gewählt, mit dem zweidimensionale Bildstrukturen beschrieben werden können. Dazu wird mit einem Diffusionsverfahren für jeden Rasterpunkt p Information über die zweidimensionale Verteilung von v berechnet (hier in Grauwerten dargestellt). Der Gradient $\mathbf{g}$ beschreibt den vorherrschenden Gradient von v in der Umgebung. Die Translation $\mathbf{r}$ beschreibt, wie weit dieser von dem Rasterpunkt p entfernt ist. Die Konfidenzen c^r und c^g beschreiben mit welcher Sicherheit die Strukturbeschreibungen Translation und Gradient ermittelt werden konnten

jedes Detektionselement vorab festgelegte Eingangssignale. Dadurch ist aber auch schon beschränkt, was ein Detektor überhaupt erkennen kann.

Schließen sich höhere Detektionsstufen an, werten Sie meist nur die detektierten Eigenschaften der direkt unterliegenden Stufen aus. Sie haben dadurch keine Möglichkeit, noch spezifisch auf einzelne Detektionsergebnisse unterer Stufen zu sein. Dies ist aber für das Erkennen vieler Objekte wesentlich, da ihre Teilrepräsentationen keine Monohierarchien bilden - also jede Teilrepräsentation nur maximal einer höheren Teilrepräsentation zugeordnet wäre. Sollen beispielsweise gelbe Postkästen erkannt werden, sollte eine untere Detektionsschicht quaderförmige Objekte aus Helligkeits- und Farbstrukturen im Bild detektieren können. Diese Detektionsstufe sollte bezüglich Farben invariant sein, weil sie aus Effizienzgründen auch noch andere quaderförmige Objekte erkennen soll. Eine höhere Detektionsstufe soll dann komplette Postkästen als gelbe, quaderförmige Objekte mit horizontaler Klappe erkennen, kann jedoch keine Spezifität mehr zum Merkmal „gelb“ bilden, da dies schon von der unteren Detektionsstufe behandelt und nicht weitergereicht wurde.

Diese Einschränkung hat das Provadeo-Verfahren nicht. Im Zuge des Lernverfahrens werden beliebig viele Schichten eingeführt, die hauptsächlich auch nur auf einige Komponenten der direkt unterliegenden Detektionsstufen spezifisch sind. Prinzipiell kann jedoch jede Schicht auf jede Komponente jeder unterliegenden Schicht spezifisch sein. Mit der Freigabe $\mathcal{E}(l)$ werden nur die Komponenten gewählt, die zur Unterscheidung der gelernten Konzepte besonders geeignet sind.

Eine Beschränkung auf möglichst wenige, effiziente Komponenten ist

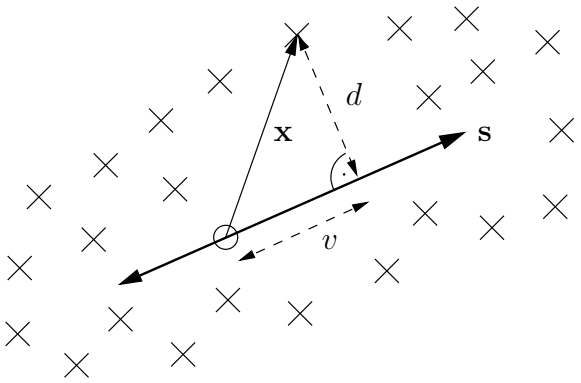


Abbildung 4.3: Die Provadero-Projektion beschreibt die Berechnung der Ausprägung v und der zugehörigen Konfidenz c. Die Spezifität $\mathbf{s}$ repräsentiert die Richtung der größten Varianz, die die Zustandsvektoren der Trainingsmuster haben (hier durch Kreuze dargestellt). Ein neuer $\mathbf{x}$ Zustandsvektor aus einem zu prüfenden Eingabebild wird auf die Spezifität projiziert. Der Wert v beschreibt, welche Ausprägung er relativ zu den bereits gelernten Zustandsvektoren hat. Der Abstand d beschreibt, wie gut er noch in der Schar der gelernten Zustandsvektoren liegt. Kleine d-Werte erzeugen eine hohe Konfidenz c.

zusätzlich wichtig. Würde stets der komplette Zustandsvektor ausgewertet, könnten sich kaum Spezifitäten für abstrakte Teilkonzepte einstellen, da die unteren Schichten dabei wie Rauschen stören würden.

4.2.2 Wie werden Repräsentationen integriert?

Jede Schicht hat eine eigene Freigabefunktion $\mathcal{E}(l)$, die beschreibt, aus welchen Komponenten der Zustandsvektor $\mathbf{x}$ zusammengestellt wird. Seine Werte bilden die Eingangssignale für die Detektionselemente. Diese berechnen anhand einer Projektion $\mathcal{P}$ die v und c Werte der Schicht (Abbildung 4.3).

Diese Werte stellen invariante und variante Anteile des der Schicht zugeordneten Teilkonzeptes dar. Die Ausprägung v berechnet sich aus der Projektion auf die Gerade der größten Varianz der bisher gelernten Zustandsvektoren. Der Wert beschreibt also, wie die Ausprägung des Eingabebildes relativ zu den Ausprägungen der Zustandsvektoren gelegen ist. Dies geschieht auf der Geraden, die durch die Varianzmaximierung große Musterabstände und damit Unterscheidbarkeit ermöglicht. Die Gerade wird durch einen Vektor - die Spezifität $\mathbf{s}$ beschrieben.

Die Konfidenz (oder auch Aktivierung) c berechnet sich aus dem Abstand, den die Projektion des Zustandsvektors senkrecht zur Geraden hat. Ist der Abstand klein, bedeutet das, dass der Zustandsvektor in einem Bereich des Merkmalsraumes liegt, in dem auch der Großteil der Punkte liegt, die den Trainingsmustern entsprechen. In diesem Fall ergibt sich eine hohe Konfidenz. Ist der Abstand groß, ergibt sich eine kleine Konfidenz.

Die Projektion ist die Funktion im Provadero-Verfahren, die eine Informationsreduktion im Sinne des in Abschnitt 3.3.3 diskutierten Charakteristika besitzt.

4.2.3 Wie werden Eigenschaften an Repräsentationen gebunden?

Die meisten Betrachtungen, die das sogenannte „Bindeproblem“ betreffen, gehen von einer expliziten Verknüpfung von symbolischen Repräsentationen aus (Wickelgren, 1969; von der Malsburg and Schneider, 1986; von der Malsburg, 1994). Im Rahmen dieses Beitrages soll eine andere Betrachtungsweise vorgeschlagen werden, die das Binden von Eigenschaften als eine Ausprägungsvariante versteht. Für jedes (Teil-)Konzept gibt es eine Ausprägung, die in beliebiger Intensität auftreten kann. Diese analoge Repräsentation kann schon für eine Kombination von symbolischen Repräsentationen stehen. Wenn unterschiedliche Eigenschaftskombinationen gebunden werden sollen und diese korreliert auftreten, kann die Ausprägung diese repräsentieren. Bei unkorrelierten Eigenschaften können mehrere repräsentierende (Teil-)Konzepte eingesetzt werden.

Die Repräsentationen von Eigenschaften parametrisieren einen eigenen Raum. So können mit wenig Dimensionen viele Kombinationen repräsentiert werden. Bei klassischen Lösungen zum Bindeproblem, wird sonst für jede Eigenschaftskombination ein separater Detektor benötigt (Wickelgren, 1969).

4.2.4 Wie fließt bekanntes Wissen mit in den Erkennungsvorgang ein?

Die wesentliche Wissenkomponente im Provadero-Verfahren ist die Spezifität $\mathbf{s}$, die für jede Schicht und damit für jedes (Teil-)Konzept gegeben ist. Sie bestimmt, für welche Eingangssignale $\mathbf{x}$ Aktivitäten erzeugt werden.

Der Erkennungsvorgang wird auch noch durch eine Rücktransformationsfunktion $\mathcal{B}$ beeinflusst. Sie versucht die Eingangssignale $\mathbf{x}$ zu normalisieren (ähnlich wie das Shifter-Circuit-Verfahren aus Abschnitt 3.3.1). Die

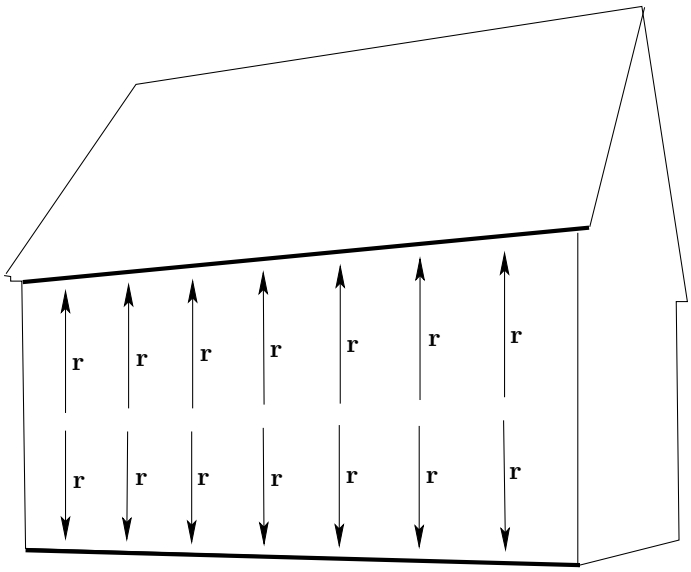


Abbildung 4.4: Perspektivische Bildverzerrungen ergeben auch ein verzerrtes Vektorfeld von Translationsvektoren. Beispielhaft werden zwei in der Realität parallel verlaufende Linien unter perspektivischer Transformation in einem Winkel aufeinander zu laufen (fett dargestellt). Die Transformation verzerrt auch das Vektorfeld der Translationsvektoren **r**. Im gezeigten Beispiel sind diese links kürzer als rechts. Diese Information kann genutzt werden, um eine Rücktransformationsfunktion so zu parametrisieren, dass eine invariante Erkennung der Hauswand möglich wird.

Parameter dieser Funktion stellen ebenfalls Wissen für den Erkennungsvorgang dar.

Die Rücktransformation soll es ermöglichen, globale Varianzen zurückzutransformieren, z.B. perspektivische Bildverzerrungen. Dazu können für die strukturbeschreibenden Vektoren **r** und **g** Winkel und Längen beeinflusst werden. Die Rücktransformation ist eine Funktion aller zuvor detektierten Ausprägungen. So ist es möglich, dass auch höhere Teilkonzepte noch eine Entzerrung der Bildgeometrie vornehmen können, die schon auf viel niedrigeren Stufen detektiert wurde.

Zur Veranschaulichung soll eine Abbildung mit einer perspektivischen Verzerrung von zwei Bildlinien betrachtet werden, die in der realen Welt parallel verlaufen. In der Abbildung laufen sie in einem bestimmten Winkel aufeinander zu. Zwischen den Linien ergibt sich auf den Pixeln ein Vektorfeld von Translationsvektoren (Abbildung 4.4).

Dieses Vektorfeld kann genutzt werden, um über die Rücktransformation

die Länge der Translationsvektoren von Folgeschichten so zu beeinflussen, dass deren Projektion so erfolgen kann, als wäre keine perspektivische Verzerrung gegeben.

Benachbarte Bildregionen benötigen typischerweise ähnliche Rücktransformationen. Diese sind aber nicht immer eindeutig aus den lokalen Bilddaten zu bestimmen. Das Provaderoverfahren stellt ein zusätzliches Diffusionsverfahren bereit, das die Parametrisierung für die Rücktransformationsfunktion interpoliert.

Die Rücktransformation ist die Funktion im Provadero-Verfahren, die eine Informationstransformation im Sinne der in Abschnitt 3.3.3 diskutierten Charakteristika besitzt. Zusammen mit der Integration von Repräsentationen ergibt sich also ein gemischtes informationsreduzierendes und -transformierendes Verfahren.

4.3 Lernen

Das Provadero-Verfahren soll eine Menge O von Konzepten unterscheiden können. Diese Konzepte erscheinen auf Abbildungen in beliebigen Varianzen. Damit die Provadero-Erkennung auch eine bisher nicht gelernte Konzeptabbildung dem richtigen Konzept zuordnen kann, muss sie zwischen Konzeptabbildungen generalisieren können. Die Provadero-Lernverfahren parametrisieren die Erkennung so, dass für Abbildungen eines Konzeptes generalisiert werden kann und Abbildungen unterschiedlicher Konzepte unterschieden werden können.

Um eine gute Generalisierung zu erreichen, wird ein Konzept o anhand von mehreren typischen Abbildungen gelernt. Dazu werden eine Menge Q^L von Trainingsmustern verwendet. Ein Trainingsmuster q^L besteht aus einem Bild ψ und einem Konzept o: $q^L = \{\psi, o\}$. Die Bilder der Trainingsmuster eines Konzeptes sollen das Konzept in anwendungstypischen Varianzen darstellen.

In jeder Lerniteration wird eine Gruppe Ω von Schichten hinzugefügt. Jeder Schicht l der Gruppe ist genau ein Konzept $o(l)$ zugeordnet. Die Schichten der Gruppe werden so eingestellt, dass sie besonders gut auf Ansichten des zugeordneten Konzeptes reagieren. Abbildung 4.5 zeigt beispielhaft für zwei Muster die Anordnung von Gruppen und Schichten.

Die Lerniterationen werden solange fortgesetzt und Schichtengruppen werden solange eingeführt, bis eine ausreichende Erkennungsleistung erreicht ist.

Beim Erkennen von Szenen wird mit dem Provadero-Verfahren eine Repräsentation erstellt, die geeignet ist, für universelle Szenenobjekte Identi-

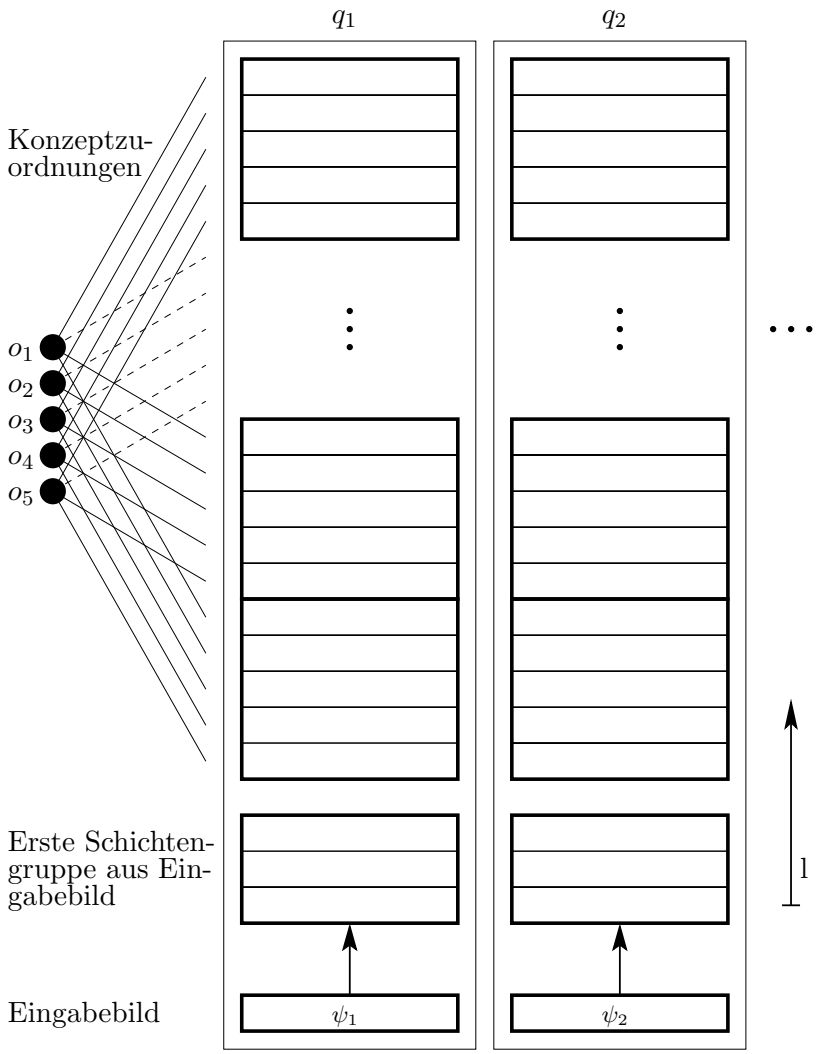


Abbildung 4.5: Jede Schichtengruppe Ω (hier fett umrandet) enthält für jedes Konzept o eine Schicht. Die unterste Gruppe enthält soviel Schichten, wie Kanäle zur Farbkodierung des Eingabebildes ψ ausgewählt wurden. Mit jeder Lerniteration wird eine Schichtengruppe neu eingeführt und parametrisiert. Jedem Muster q ist ein eigener Stapel von Schichtengruppen zugeordnet. Die Assoziationsmodule einer Schicht l werden durch die Lernverfahren für alle Stapel identisch parametrisiert. In Abhängigkeit des Eingabebildes und der Assoziation ergeben sich dann aber unterschiedlich aktivierte (Teil-)Konzepte auf der Schicht.

fikationen und Klassifikationen vorzunehmen.

Damit diese Verarbeitung möglich wird, stellt das Lernverfahren vorab die Parameter für die Freigabe, die Projektion und die Rücktransformation ein.

Für die Freigabefunktion werden die Schichten ausgewählt, deren mittlere Aktivierung am besten zur Abgrenzung vom (Teil-)konzept der lernenden Schicht gegenüber den restlichen (Teil-)konzepten geeignet sind.

Für die Projektion wird die Spezifität **s** so eingestellt, das sie entlang der größten Varianz der zu lernenden Zustandsvektoren liegt.

Die Rücktransformation wird so eingestellt, dass die folgenden Schichten möglichst invariante Signale eines Konzeptes erhalten. Dabei muss eine Unterscheidbarkeit der unterschiedlichen Konzepte erhalten bleiben.

4.4 Varianzen

4.4.1 Wie funktioniert Erkennen unter klassischen Varianzen?

Alle Detektionselemente sind auf den Rasterpositionen einer Schicht gleich parametrisiert. Dadurch ist intrinsisch eine Verschiebunginvarianz gegeben. Die anderen klassischen Varianzen sollen in erster Linie durch die Rücktransformationsfunktion aufgelöst werden.

4.4.2 Wie funktioniert Erkennen unter abstrakten Varianzen?

Wie schon in Abschnitt 2.2.6 beschrieben, sind abstrakte Varianzen wohl nur aufzufangen, wenn viele Detektionsstufen verwendet werden. Die optimale Anzahl hängt von der Anzahl der Konzepte und deren typischen Ausprägungen in der Anwendung ab. Sind schon einfache Eigenschaften der Konzepte in den Abbildungen disjunkt, genügen wenige Detektionsstufen. Dies wird zum Beispiel beim Roboterfußball ausgenutzt, wo in nur einer Stufe eine Objekterkennung mittels Farbsegmentierung möglich ist. Treten jedoch Teilkonzepte immer wieder in anderen Ausprägungen auf, werden mehrere Stufen benötigt. Beispielsweise haben Telefone immer eine Wähleinrichtung - diese kann als Tastatur oder Wählscheibe ausgeführt sein. Ein einzelner Taster einer Telefontastatur kann eine runde oder eckige Form haben, etc. Sollen zum Beispiel Telefone, Laptops und Aktenordner voneinander unterschieden werden, sind Tastaturen und Taster effiziente

Teilkonzepte deren invariante Repräsentation wertvoll für folgende Detektionsstufen ist.

Es ist schwierig, die optimale Anzahl von Detektorstufen analytisch zu bestimmen, da das insbesondere von der Repräsentation der zweidimensionalen Struktur von Teilkonzepten und der Detektoreigenschaften abhängt. Für die Provadero-Architektur wird die Anzahl der Detektionsstufen empirisch bestimmt. Es werden iterativ solange Detektionsstufen (Schichten) hinzugefügt, bis die Erkennung ausreichend gut unterscheiden und generalisieren kann.

4.5 Zusammenfassung und Bewertung

Das Provadero-Verfahren bietet Lösungen für wichtige Fragestellungen des Bildverstehens an. Es wird vorgeschlagen, Bildvarianzen universell zu verarbeiten. Das ist wahrscheinlich auch eine wesentliche biologische Eigenschaft. Es wird auch vorgeschlagen, Varianzen explizit zu repräsentieren. Dies wurde in Modellierungen bisher nicht umfassend eingesetzt.

Zur Umsetzung von Verarbeitungsfunktionen wurden Detektionselemente und Module wie Freigabe, Rücktransformation, Projektion und Lernen beschrieben. Eine mögliche Modulrealisierung wird im nächsten Kapitel beschrieben. Andere Realisierungen sind denkbar und können später im Rahmen weiterer Forschungsaktivitäten entwickelt werden.

Kapitel 5

Provadero-Realisierung

5.1 Assoziation

Das Provadero-Verfahren repräsentiert den Inhalt einer Bildszene durch die Aktivierung von Teilrepräsentationen. Diese sind in Schichten l angeordnet; wobei jeder Schicht eine Teilrepräsentation (ein Teilkonzept) $o(l)$ zugeordnet ist. Zusammen bilden die Schichten einen Stapel L. Die Funktion $h(l)$ liefert die Höhe einer Schicht im Stapel. Die unterste Schicht l^0 hat den Höhenwert $h(l^0) = 0$. Die Funktion $\lambda(h)$ liefert die Schicht auf der Höhe ($h \in \mathbb{N}$).

Die unterste Schicht im Stapel repräsentiert das Eingabebild. Bei Mehrkanal-Eingabebildern können auch die nächsthöheren Schichten mit zur Repräsentation des Eingabebildes verwendet werden.

Das Eingabebild besteht aus I horizontalen und J vertikalen Pixeln. Die Gesamtanzahl P der Pixel berechnet sich aus $P = IJ$.

Auf jeder Schicht liegt ein Raster mit Positionen p. Die Raster entsprechen mit IJ Positionen der Auflösung des Eingabebildes.

Die Aktivierung (oder auch Konfidenz) c an einem Rasterpunkt beschreibt, wie wahrscheinlich das Teilkonzept an der Position p des Rasterpunktes im Eingabebild gegeben ist und v beschreibt in welcher Ausprägung das Teilkonzept gegeben ist.

Aus diesen Werten werden dann mit einem Diffusionsverfahren $\mathcal{H}$ Beschreibungen über die zweidimensionale Struktur des Teilkonzeptes um p erstellt. Das sind die Werte $\mathbf{r}$, $\mathbf{g}$ und deren Konfidenzen c^r, c^g. Ein anderes Diffusionsverfahren $\mathcal{F}$ interpoliert die v Werte in Bereichen mit geringer c Konfidenz und erstellt daraus ein neues Wertepaar v^f und c^f. Dies Wertepaar steuert die in Abschnitt 4.2.4 beschriebene Rücktransformation auf höheren Schichten.

Die Aktivierung und die Ausprägung werden durch die Projektion be-

rechnet. Dazu werden die Strukturbeschreibungen $\mathbf{r}$, $\mathbf{g}$, c^r und c^g der unterliegenden Schichten ausgewertet. Zuvor erfolgt noch die Rücktransformation der Strukturbeschreibungen. Damit soll erreicht werden, dass globale und regionale Bildverzerrungen, die im Eingabebild z.B. durch perspektivische Projektion gegeben sind, zurückgerechnet werden können. Das Ergebnis dieser Berechnung ist der Zustandsvektor $\mathbf{x}^s$ mit der zugehörigen Konfidenz $\mathbf{c}^s$. Abbildung 5.1 zeigt die wesentlichen Elemente von Assoziation und Rücktransformation.

Die Provadero-Assoziation $\mathcal{A}$ umfasst sowohl das Überprüfen, ob eine bestimmte Kombination von Teilkonzepten in einer bestimmten Anordnung im Eingabebild gegeben ist, als auch die Rücktransformation einer Bildverzerrung, die diese Kombination von Teilkonzepten betrifft. Die einzelnen Teilschritte der Assoziation sind in Abbildung 5.2 veranschaulicht.

Durch jede Lerniteration wird eine neue Schichtengruppe Ω eingeführt. Die Assoziationsteilschritte werden für jede Schichtengruppe Ω durchgeführt. Zunächst werden die Diffusionen $\mathcal{H}(\Omega)$ und $\mathcal{F}(\Omega)$ ausgeführt. Danach folgen Freigabe $\mathcal{E}(\Omega)$, Rücktransformation $\mathcal{B}(\Omega)$, Skalierung $\mathcal{S}(\Omega)$, Projektion $\mathcal{P}(\Omega)$ und die Normierung $\mathcal{N}(\Omega)$. Diese Schritte beschreiben, wie für eine Schichtengruppe Konfidenzen, Ausprägungen, Strukturbeschreibungen und deren Konfidenzen berechnet werden. Dies geschieht in Abhängigkeit des Eingabebildes und der schon assoziierten Werte unterliegender Schichten.

Die ersten vier Schritte beschreiben die Zusammenstellung des Zustandsvektors $\mathbf{x}^s$ und der zugehörigen Konfidenz $\mathbf{c}^s$. Beide werden für den wesentlichen Assoziationsschritt: die Projektion verwendet. Der Zustandsvektor und die zugehörige Konfidenz werden dabei für jeden Rasterpunkt der assoziierenden Schicht aus Strukturbeschreibungen unterliegender Schichten berechnet. Abschließend wird $\tilde{c}$ zu c normiert. Diese Anpassung richtet sich nach der globalen Aktivität in der Schichtengruppe und stellt sicher, das deren mittlere Aktivität immer auf dem gleichen Niveau gehalten wird.

Die Assoziation wird für Trainingsmuster q^L und Testmuster q^T gleichermaßen durchgeführt. Bis auf die Normierung läuft die Assoziation für jedes Muster q unabhängig von anderen Mustern, daher wird für die Darstellung der anderen Assoziationschritte der Parameter q weggelassen.

Für die folgenden Formeldarstellungen sollen noch folgende Formatierungskonventionen gelten: Akzente und hochgestellte Buchstaben sind Bezeichner. Tiefgestellte Zeichen sind Indizes. Fettgeschriebene Variablen sind Vektoren. Großschreibung steht für Mengen oder Anzahl von Elementen einer Menge. Kalligraphische Buchstaben beschreiben Funktionen.

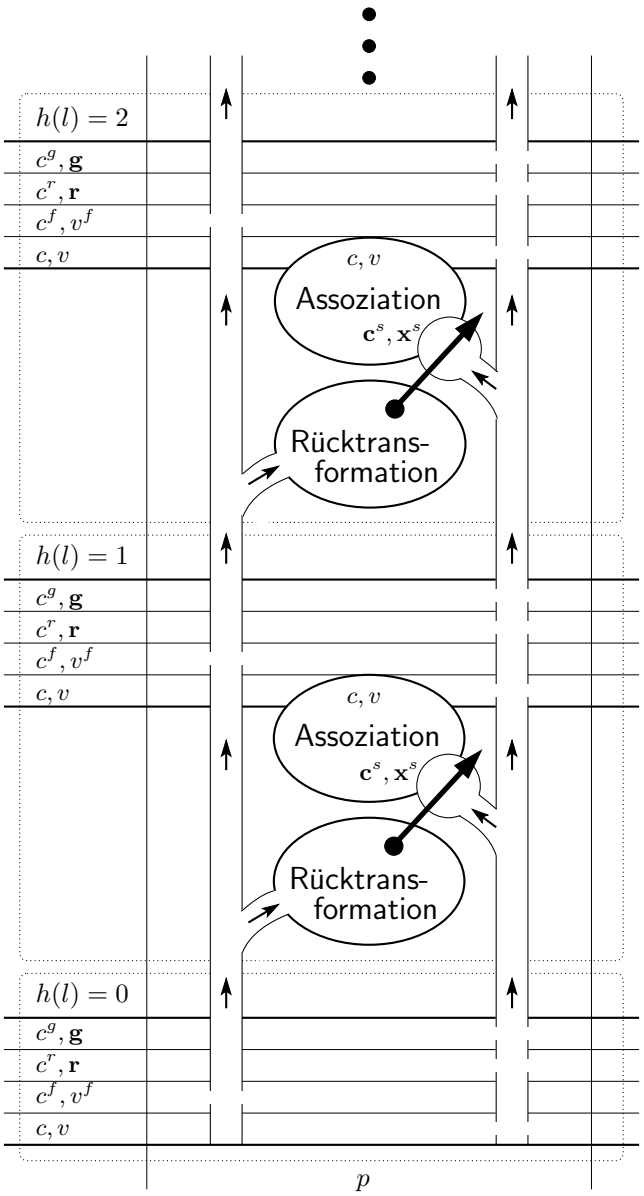


Abbildung 5.1: Provadero-Assoziation an einer Pixelposition p. Es sind drei Schichten $h(l) = 0..2$ abgebildet. Die erste Schicht ergibt sich aus dem Eingabebild (bei mehrkanaligen Eingabebildern können dies auch mehrere Schichten sein). Für alle Folgeschichten wird eine Assoziation duchgeführt, die von den Strukturbeschreibungen $v^f, \mathbf{r}, \mathbf{g}$ mit deren Konfidenzen aus unterliegenden Schichten abhängt. Diese Werte werden noch durch eine Rücktransformation verändert, die von den Ausprägungen v^f und Konfidenzen c^f unterliegender Schichten abhängt. Die Assoziation erzeugt aus den rücktransformierten Werten $\mathbf{c}^s, \mathbf{x}^s$ neue Werte c, v für die nächsthöhere Schicht. Aus diesen Werten werden dann mit einem Diffusionsverfahren die Strukturbeschreibungen der Schicht erzeugt.

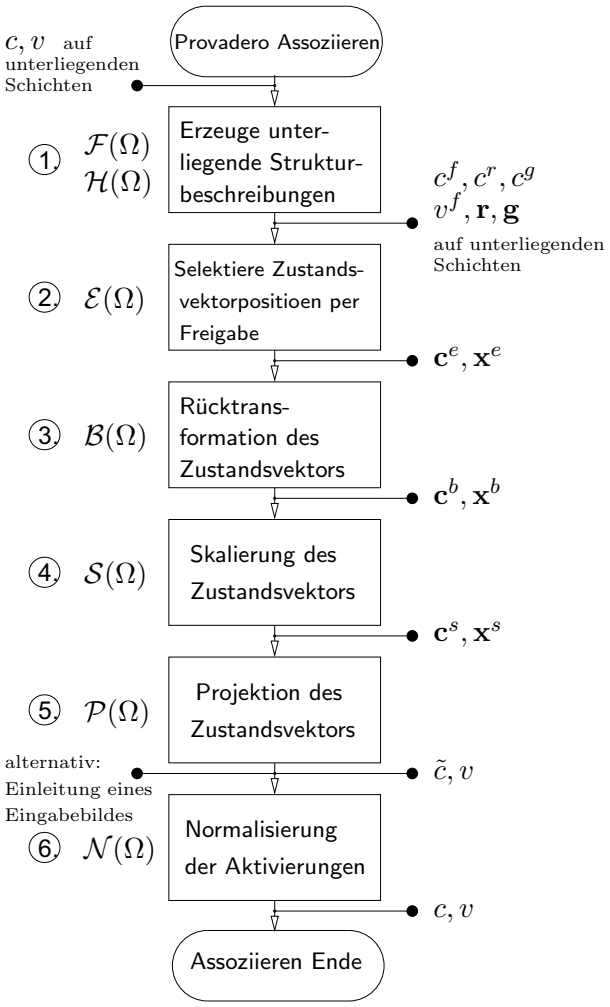


Abbildung 5.2: Flussdiagramm zur Provadero-Assoziation. Das Verfahren wird für alle Schichten angewendet. Die Bearbeitung startet auf der untersten Schicht und wird dann jeweils für die nächst höhere Schicht durchgeführt. Zunächst werden auf unterliegenden Schichten Strukturbeschreibungen mit einem Diffusionsverfahren erstellt (1). Danach wird der Zustandsvektor für die assoziierende Schicht zusammengestellt (2). Dabei finden nur Komponenten Verwendung, die einen relevanten Beitrag zur Assoziation leisten können, was durch eine Freigabe gesteuert wird. Als nächstes wird eine Rücktransformation vorgenommen, die diese Schicht betreffende Varianzen eliminieren soll (3). Nachdem der Zustandsvektor skaliert wurde (4), kann der wesentliche Assoziationsschritt - die Projektion durchgeführt werden (5). Dabei werden die Aktivierung und die Ausprägung für die Folgeschicht bestimmt. Abschließend werden die berechneten Aktivierungen für die assoziierende Schichtengruppe normiert (6). Für die untersten Schichten, die Informationen der Eingabebilder bereitstellen, ist das der einzige Schritt.

5.1.1 Diffusionsverfahren

Jedes Verfahren zur Bildanalyse muss auf irgendeine Art und Weise Informationen über den visuellen Kontext also die strukturelle Ausprägung eines zweidimensionalen Bereichs auswerten können. Diese Notwendigkeit bringt erhebliche Schwierigkeiten mit sich, da die zweidimensionalen Strukturen Transformationen, wie z.B. perspektivischen Varianzen unterliegen. Viele Verfahren zum Konzepterkennen starten in einem ersten Bildverarbeitungsschritt mit einer Detektion einfacher Strukturen und benutzen dazu Filter, die eine Faltungsoperation einer Filtermaske auf dem Bild durchführen. Diese Maske gibt eine feste zweidimensionale Gewichtung der Umgebung vor und ist damit im Allgemeinen nicht mehr invariant gegenüber z.B. Skalierungen oder allgemeinen perspektivischen Varianzen.

Das Provadero-Verfahren verwendet keine derartig starre Vorverarbeitung. Auf jeder Schicht wird eine Kontextanalyse mit einem Diffusionsverfahren durchgeführt, das einfache Strukturbeschreibungen iterativ erstellt. Es basiert auf dem von Teichert and Malaka (2006) vorgestellten Diffusionsverfahren und wird hier in einer erweiterten Form für die Diffusion $\mathcal{H}$ der Strukturwerte und die Diffusion $\mathcal{F}$zur Interpolation der Ausprägungswerte verwendet.

Nachdem die Projektion auf einer Schicht für jeden Rasterpunkt eine Aktivierung c und eine Ausprägung v berechnet hat, werden Informationen über deren Veränderungen gegenüber benachbarten Rasterpunkten über die Schicht diffundiert.

Mit der $\mathcal{F}$ Funktion wird die Ausprägung auch für Gebiete bestimmt, die eine geringe Konfidenz haben. Die Wirkungsweise entspricht in etwa einer Interpolation. Die interpolierten Werte sind dann die Aktivierung c^f und die Ausprägung v^f. Sie werden zur Steuerung der Rücktransformation verwendet.

Die $\mathcal{H}$ Funktion erstellt für jede Position p zwei Vektoren: der vorherrschende Gradient in der Umgebung $\mathbf{g}$ und die Translation $\mathbf{r}$, die beschreibt, wo sich $\mathbf{g}$ relativ zu p befindet. Die Konfidenzen c^r und c^g beschreiben, wie verlässlich die Translation und der Gradient ermittelt wurden. Abbildung 4.2 zeigt beispielhaft, wie so eine Repräsentation aussehen kann. Alle Werte $\mathbf{g}$, $\mathbf{r}$, c^r und c^g werden direkt an der Position p repräsentiert.

Diese Kontextrepräsentation ist relativ primitiv. Werden jedoch mehrere Schichten betrachtet, können komplexe Inhalte repräsentiert werden.

Die Diffusionsverfahren lassen sich abstrakt auch gut anhand einer Simulation von fließenden farbigen Flüssigkeiten veranschaulichen. Dazu wird angenommen, dass auf jedem Pixel eine Flüssigkeitssäule steht und sich ihre initiale Höhe nach der Aktivierung c richtet. Weiter wird für die $\mathcal{F}$ Funktion

angenommen, dass die Farbe durch die Ausprägung v gegeben ist. In einem Simulationsschritt werden Bereiche mit geringer Füllhöhe mit der Farbe aus angrenzenden Bereichen mit hoher Füllhöhe aufgefüllt. Der Austausch erfolgt nur in Abhängigkeit der im Simulationsschritt gegebenen Differenzen. Ein Farbaustausch bei gleicher Füllhöhe findet nicht statt (bei diffusionsoffenen Trennwänden wäre das in der Realität der Fall). Das Wassermodell kann man sich also so vorstellen, dass alle (Pixel-)Behälter voneinander getrennt sind und in jedem Simulationsschritt Differenzen in der direkten Umgebung umgefüllt werden.

Bei der $\mathcal{H}$ Funktion sollen Differenzen in der umliegenden Struktur untersucht werden. Für die Flüssikeitsanalogie kann man die Farbe nun als Richtung deuten, in der ein lokaler Gradient gegeben ist. Die Stärke des Gradienten entspräche dann der Farbsättigung. In einer Diffusionssimulation mit Schwerkraft werden nun wieder Bereiche mit geringer Füllhöhe von angrenzenden mit hoher Füllhöhe überschwemmt. Dabei breitet sich die Farbe des stärksten Gradienten in der Umgebung aus. Gleichzeitig wird gemessen, wie weit die Flüssigkeit geflossen ist und aus welcher Richtung sie gekommen ist. Das ergibt den Translationsvektor. Dessen Aktivierung bleibt hoch, wenn es auf dieser Flusstrecke zu wenig Farbvermischungen gekommen ist.

Diffusionsverfahren werden erfolgreich zur Entrauschung von Bildern eingesetzt (z.B. Didas et al., 2009). Die Verwendung im Provadero-Verfahren zur Erstellung von Strukturbeschreibungen ist jedoch ein neues Anwendungsfeld.

Rastereigenschaften

Die im Folgenden vorgestellten Diffusionsverfahren definieren den Informationsaustausch zwischen Zellen, die in einem Raster angeordnet sind. Es sind drei Rastertopologien möglich, wenn die Zellen konvex und flächenfüllend und das Raster regelmäßig sein sollen. D.h. die Abstände zu den Zellschwerpunkten innerhalb einer Nachbarschaftsklasse sind identisch. Abbildung 5.3 zeigt Abstände für gleichseitige Dreiecke, Quadrate und Hexagone.

Der Abstandsfaktor ς drückt aus, wie weit zwei Zellmittelpunkte im Verhältnis zu einer direkten Nachbarschaft voneinander entfernt sind. Er findet Verwendung, wenn Informationen aus der Umgebung aufsummiert und dabei in Abhängigkeit ihres Abstandes gewichtet werden müssen. Dabei wird die entsprechende Richtung mit einem tiefgestellten Index ausgedrückt. Wenn z.B. über x-Werte aller Nachbarschaftszellen gewichtet summiert werden soll, kann das einfach mit $\sum_a \varsigma_a x(p_a)$ ausgedrückt werden.

Nachbarschaft:		direkt		sekundär		tertiär	
Topologie:	Kanten-länge	#	ς	#	ς	#	ς
gleichs. Dreiecke	$\sqrt{3}$	3	1	6	$\sqrt{3}$	3	2
Quadrate	1	4	1	4	$\sqrt{2}$	-	-
Hexagone	$1/\sqrt{3}$	6	1	-	-	-	-

Abbildung 5.3: Abstandsfaktoren von Rasterzellen in Abhängigkeit der Rastertopologie. Direkte Nachbarzellen teilen die gleichen Kanten. Sekundäre und tertiäre teilen nur Eckpunkte. Die sekundären Nachbarschaftszellen haben dabei den geringeren Abstand. Die Tabelle zeigt die Anzahl # der jeweiligen Nachbarschaftszellen und deren Abstand ς zur betrachteten Zelle. Gemessen wird dieser zwischen den Flächenschwerpunkten der Zellen. Alle Abstände sind auf den Abstand der direkten Nachbarn normiert. Die Kantenlänge der Zellen und die Abstände zu den sekundären und tertiären Nachbarn ergeben sich also mit der Einheit des Abstandes der direkten Nachbarn.

Von einer Zelle an der Position p kann die Nachbarschaftszelle in Richtung a einfach über p_a referenziert werden.

Damit gleich auch eine Normierung auf die Summe der Abstandsfaktoren vorgenommen werden kann, wird ein entsprechend erweiterter Faktor σ eingeführt:

$$\sigma_a = \frac{\varsigma_a}{\sum_{a'} \varsigma_{a'}}. \tag{5.1}$$

Wird ein Wert über eine Richtungssumme berechnet, wie beispielsweise $y = c \sum_a \sigma_a \, x(p_a)$, so bezeichnet ein tiefgestellter Index an dieser Variablen den Summenanteil in der entsprechenden Richtung: $y_a = c \, \sigma_a \, x(p_a)$.

Das Provadero-Verfahren verwendet nur Raster mit quadratischen Zellen, da sich für aktuelle Rechnerstrukturen erhebliche Implementierungsvorteile ergeben. Es gibt also nur direkte und sekundäre Nachbarn.

Abbildung 5.4 zeigt die zugehörigen Nachbarschaften. Für die direkten Nachbarn die vierfache Nachbarschaft, die im Folgenden mit a indexiert wird und für die sekundären Nachbarn die achtfache Nachbarschaft, die im Folgenden mit b indexiert wird.

5.1.2 Diffusionsverfahren zur Interpolation

Die c Werte sind nach einer Projektion typischerweise nur spärlich aktiviert. Für die in Abschnitt 4.2.4 beschriebene Rücktransformation werden jedoch auch Werte in Bereichen benötigt, die nicht durch die Assoziation aktiviert

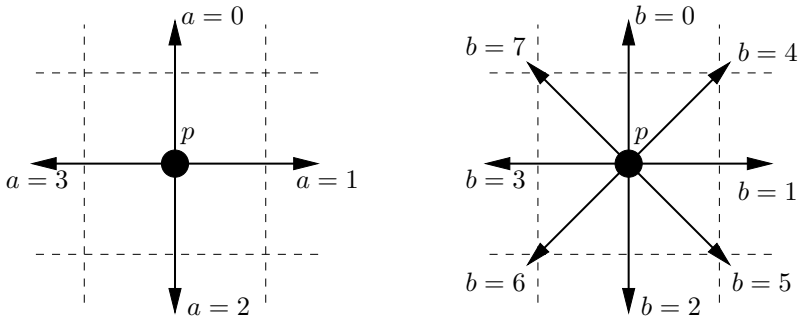


Abbildung 5.4: Nachbarschaftsrichtungen für einen Rasterpunkt an der Position p für vier Nachbarn (a) und acht Nachbarn (b). Die gestrichelten Linien zeigen Grenzen zwischen Rasterzellen.

werden. Dafür werden die Werte c^f und v^f eingeführt, die per Diffusion eine Art Interpolation realisieren.

Obwohl die Rücktransformation in diesem Beitrag nicht evaluiert wird, soll das dazu entwickelte Diffusionsverfahren zur Interpolation $\mathcal{F}$ hier vorgestellt werden. Das Diffusionsverfahren für Strukturwerte $\mathcal{H}$ funktioniert nach gleichen Grundprinzipien, ist aber komplexer und dadurch unübersichtlicher. Die Darstellung der Funktionsweise der $\mathcal{F}$-Diffusion soll mit den Diffusionsverfahren auf einer einfacheren Ebene vertraut machen.

Die Diffusion zur Interpolation $\mathcal{F}(\Omega)$ berechnet die Werte c^f und v^f für eine Schichtengruppe Ω. Die Berechnung erfolgt iterativ über T^f Schritte, die mit t indiziert werden. Die Diffusion erstreckt sich über alle Positionen p auf den Schichten l, die zur aktuell betrachteten Schichtengruppe gehören. Die Variablen starten mit Nullwerten:

$$\left.\begin{array}{rcl} c^f(l,p,t=0) & = & 0 \\ v^f(l,p,t=0) & = & 0 \end{array}\right\} \quad \forall\, p,l \in \Omega \tag{5.2}$$

Grundlage für die Diffusion ist der Austausch der Konfidenzwerte. Die Berechnung der Ausprägungswerte richtet sich nach den ausgetauschten Konfidenzanteilen und den zugehörigen Ausprägungen. Abbildung 5.5 zeigt für einen Rasterausschnitt, welche Werte relativ zur Position p eine Rolle bei der $\mathcal{F}$-Diffusion spielen.

Die Konfidenz im Folgeschritt (t+1) berechnet sich aus einem lokalen Anteil $\dot{c}^f$ und einem Anteil $\mathring{c}^f$ mit aufsummierten Werten aus der Umgebung:

$$c^f(l,p,t+1) \;=\; \dot{c}^f(l,p,t) + \mathring{c}^f(l,p,t) \tag{5.3}$$

Der Informationsaustausch wirkt nur auf die unmittelbare Umgebung. Werden jedoch mehrere Iterationen durchgeführt, können Aktivierungen über

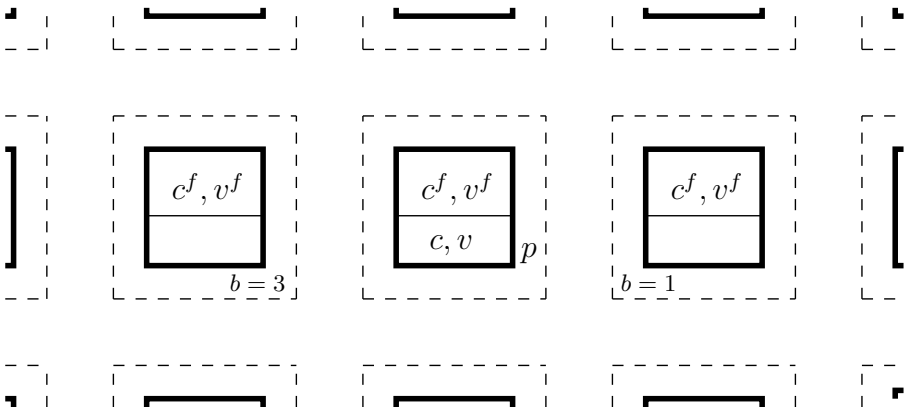


Abbildung 5.5: Repräsentationsschema für die Diffusion von c^f und v^f an einem Rasterpunkt p. Die Werte c und v, die mit der Assoziation berechnet wurden, liefern einen lokalen Anteil. Die Werte c^f und v^f, die aus der Umgebung Anteile zur Diffusion beitragen, sind in diesem Ausschnitt beispielhaft für zwei benachbarte Rasterpositionen gezeigt.

größere Strecken weitergereicht werden. Die Anteile berechnen sich wie folgt:

$$
\begin{aligned}
& c^f(l,p,t+1) \\
&= \underbrace{\frac{c(l,p)}{c(l,p)+\sum_b \sigma_b c^f(l,p_b,t)}\, c(l,p)}_{\dot{c}^f(l,p,t)} + \qquad (5.4) \\
&\qquad \underbrace{\frac{\sum_b \sigma_b c^f(l,p_b,t)}{c(l,p)+\sum_b \sigma_b c^f(l,p_b,t)} \sum_b \sigma_b c^f(l,p_b,t)}_{\mathring{c}^f(l,p,t)} \\
&= \frac{1}{c(l,p)+\sum_b \sigma_b c^f(l,p_b,t)} \left(\Big(c(l,p)\Big)^2 + \Big(\sum_b \sigma_b c^f(l,p_b,t)\Big)^2 \right) \qquad (5.5)
\end{aligned}
$$

Wesentlicher Bestandteil des lokalen Anteils $\dot{c}^f$ ist die Konfidenz c. Für den Anteil aus der Umgebung $\mathring{c}^f$ ist es die mit den Abstandsfaktoren gewichtete Summe der kumulierenden c^f benachbarter Zellen. Welchen Beitrag diese Anteile auf die Folgekonfidenz haben, bestimmt ihre eigene Größe. Mit dieser Gewichtung wird gewährleistet, dass die Folgekonfidenz von den Anteilen mit höheren Werten bestimmt wird. Diese Gewichtung ist so normiert, dass sie in der Summe eins ergibt. Abbildung 5.6 zeigt, wie die Zuweisungen als Austausch von Flüssigkeiten visualisiert werden können.

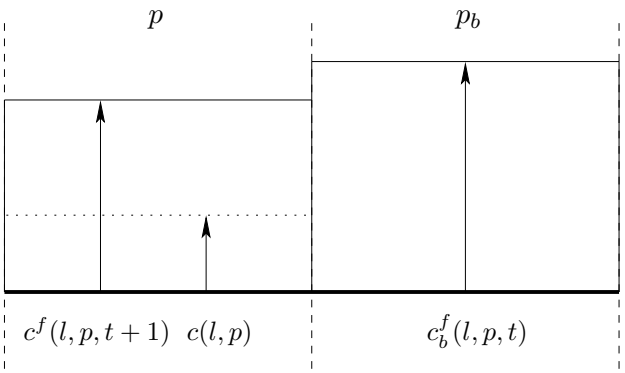


Abbildung 5.6: Betrachtet man die $\mathcal{F}$-Diffusion als Austausch von Flüssigkeiten, so kann man sich für jede Zelle einen Behälter vorstellen, in dem die Konfidenzen analog der Füllhöhe sind. Die $\mathcal{F}$-Diffusion berechnet die Füllhöhe $c^f(t+1)$ aus der (Eingabe-)Konfidenz c und den Konfidenzen aus der Nachbarschaft $c_b^f(t)$. In diesem Beipiel würde mehr Konfidenz aus der Nachbarzelle in die Zelle an der Position p „fließen“. Die neue Füllhöhe ist dabei nicht direkt abhängig vom aktuellen Wert $c^f(t)$. Sie wird in jedem Iterationsschritt neu berechnet. Allerdings wirkt sie in jeder Iteration auch auf Ihre Umgebung. Dadurch ergibt sich ein regionales „Gedächtnis“. Die Ausprägungen v der Zellen können als Farben der Flüssigkeitssäulen in den Zellen gedeutet werden. Die Farbe im Folgeschritt richtet sich nach den ausgetauschten Konfidenzanteilen.

Zur Berechnung der Ausprägung im Folgeschritt, werden wieder ein lokaler Anteil $\dot{v}^f$ und ein Anteil $\mathring{v}^f$ aus der Umgebung betrachtet:

$$v^f(l,p,t+1) \;=\; \dot{v}^f(l,p,t) + \mathring{v}^f(l,p,t). \tag{5.6}$$

Die Ausprägung v ist wesentlicher Bestandteil des lokalen Anteils. Für den Anteil aus der Umgebung ist es v_b^f. Die Gewichtung erfolgt anhand der zugehörigen Konfidenzanteile aus (5.3). Damit sich diese Anteile auf eins aufsummieren, wird mit der Folgekonfidenz normiert:

$$\begin{aligned} & v^f(l,p,t+1) \\ &= \underbrace{\sum_b \sigma_b \frac{\dot{c}_b^f(l,p,t)}{c^f(l,p,t+1)}\, v(l,p)}_{\dot{v}^f(l,p,t)} + \underbrace{\sum_b \sigma_b \frac{\mathring{c}_b^f(l,p,t)}{c^f(l,p,t+1)}\, v^f(l,p_b,t)}_{\mathring{v}^f(l,p,t)} \qquad (5.7) \\ &= \frac{1}{c^f(l,p,t+1)} \Big(\dot{c}^f(l,p,t)\, v(l,p) + \sum_b \sigma_b \mathring{c}_b^f(l,p,t)\, v^f(l,p_b,t) \Big). \qquad (5.8) \end{aligned}$$

Nachdem die T^f Iterationen durchgeführt wurden, stehen die Werte c^f und v^f für die folgenden Verarbeitungsschritte zur Verfügung:

$$c^f(l,p) = c^f(l,p,t=T^f), \tag{5.9}$$
$$v^f(l,p) = v^f(l,p,t=T^f). \tag{5.10}$$

Eine sinnvolle Anzahl der Iterationsschritte hängt von der Bildgröße und vom Bildinhalt ab. Ein sicherer Wert ist die Anzahl der Pixel in der größten Bilddimension.

Das Verfahren füllt Regionen mit geringer Konfidenz mit Ausprägungsinformation aus der Umgebung. Dadurch bleiben aktivierte Strukturen erhalten und werden auch bei einer hohen Iterationsanzahl nicht „verwaschen“ wie es sonst bei Interpolationsverfahren typisch ist.

Das Diffusionsverfahren ist parameterfrei. Eine implizite Parametrisierung ergibt sich jedoch aus der Gleichgewichtung des lokalen Anteils und des Anteils aus der Umgebung.

5.1.3 Diffusionsverfahren für Strukturwerte

Mit dem Diffusionsverfahren $\mathcal{H}(\Omega)$ wird eine Beschreibung über die Struktur der Ausprägungswerte v in der Umgebung erstellt. Dies erfolgt für jede Position p auf jeder Schicht l der Schichtengruppe Ω. Der Gradient $\mathbf{g}$ beschreibt den dominanten Gradienten von v in der Bildumgebung von p und die Translation $\mathbf{r}$ beschreibt, wie der Gradient relativ zu p gelegen ist (Abbildung 4.2). Für beide Vektoren beschreiben zugehörige Konfidenzen c^g, c^r wie sicher diese Repräsentationen sind.

Die Diffusion wird für T^h Iterationen durchgeführt und dabei mit t indiziert. Zu Beginn sind alle Werte null gesetzt:

$$\left.\begin{array}{rcl} c^g(l,p,t=0) &=& 0 \\ c^r(l,p,t=0) &=& 0 \\ \mathbf{g}(l,p,t=0) &=& 0 \\ \mathbf{r}(l,p,t=0) &=& 0 \end{array}\right\} \quad \forall\, p,l \in \Omega \tag{5.11}$$

Wie bei dem Diffusionsverfahren zur Interpolation werden auch bei diesem Verfahren Informationen über umliegende Konfidenzen ausgetauscht. Der Gradient berechnet sich aus der Differenz der Ausprägung benachbarter Zellen. Die Weiterleitung des Gradienten hängt von den zugehörigen Konfidenzen ab. Die Translation berücksichtigt, wie geradlinig die Weiterleitung des Gradienten erfolgt ist.

Zur Berechnung der Folgewerte werden bei diesem Diffusionsverfahren jeweils drei Anteile berücksichtigt:

$$\begin{array}{rcl} c^g(l,p,t+1) & = & \dot{c}^g(l,p,t) + \hat{c}^g(l,p,t) + \mathring{c}^g(l,p,t), \\ c^r(l,p,t+1) & = & \dot{c}^r(l,p,t) + \hat{c}^r(l,p,t) + \mathring{c}^r(l,p,t), \\ \mathbf{g}(l,p,t+1) & = & \dot{\mathbf{g}}(l,p,t) + \hat{\mathbf{g}}(l,p,t) + \mathring{\mathbf{g}}(l,p,t), \\ \mathbf{r}(l,p,t+1) & = & \dot{\mathbf{r}}(l,p,t) + \hat{\mathbf{r}}(l,p,t) + \mathring{\mathbf{r}}(l,p,t). \end{array} \tag{5.12}$$

Die mit einem Punkt markierten Werte $\dot{c}^g$, $\dot{c}^r$, $\dot{\mathbf{g}}$ und $\dot{\mathbf{r}}$ stellen lokale Anteile dar. Sie kumulieren ausgetauschte Anteile über die Iterationen. Die mit einem Zirkumflex markierten Werte $\hat{c}^g$, $\hat{c}^r$, $\hat{\mathbf{g}}$ und $\hat{\mathbf{r}}$ stellen Werte auf Zellengrenzen dar. Sie beinhalten Differenzen von Ausprägungen benachbarter Zellen als Quellen für die Flussanteile. Mit einem Kreis markierte Werte $\mathring{c}^g$, $\mathring{c}^r$, $\mathring{\mathbf{g}}$ und $\mathring{\mathbf{r}}$ stellen Anteile aus der Umgebung dar.

Die Gradientenkonfidenz berechnet sich in der Folgeiteration zu:

$$\begin{array}{rcl} c^g(l,p,t+1) & = & \dot{c}^g(l,p,t) + \hat{c}^g(l,p,t) + \mathring{c}^g(l,p,t) \\ & = & \underbrace{\sum_b \sigma_b \dot{c}^u_b(l,p,t)}_{\dot{c}^g(l,p,t)} + \underbrace{\sum_b \sigma_b \hat{c}^u_b(l,p,t)}_{\hat{c}^g(l,p,t)} + \underbrace{\sum_b \sigma_b \mathring{c}^u_b(l,p,t)}_{\mathring{c}^g(l,p,t)} \end{array} \tag{5.13}$$

Die Anteile $\dot{c}^u$, $\hat{c}^u$ und $\mathring{c}^u$ berechnen sich über einige Zwischenschritte, die im Folgenden beschrieben werden.

Grundlegender Berechnungsschritt dieses Verfahrens ist die Differenzbildung zwischen Ausprägungen benachbarter Zellen. Der Gradient $\mathbf{g}^\delta_b$ beschreibt die Differenz der Ausprägung einer Zelle an der Position p in Bezug zu der Ausprägung ihrer Nachbarzelle in Richtung b:

$$\mathbf{g}^\delta_b(l,p) = \big(v(l,p_b) - v(l,p)\big)\,\mathbf{e}_b. \tag{5.14}$$

Dabei beschreibt $\mathbf{e}_b$ einen Einheitsvektor in Richtung b.

Die zu $\mathbf{g}^\delta_b$ gehörige Konfidenz c^δ_b bestimmt sich aus der geringeren Konfidenz der beiden Zellen, da die Differenzbildung für g^δ_b nie eine höhere Konfidenz haben kann, als die der Subtraktionskomponenten.

$$c^\delta_b(l,p) = \min\big(c(l,p_b), c(l,p)\big). \tag{5.15}$$

Auf den Schichten der Eingabebildkanäle hat die Konfidenz c^δ stets einen konstanten positiven Wert. Würden diese Konfidenzen direkt zur Bestimmung der in der Diffusion ausgetauschten Anteile verwendet werden, würde kein Austausch vorgenommen. Es soll aber auch schon auf dem Eingabebild eine Strukturanalyse möglich sein. Daher werden die Werte von c^δ und

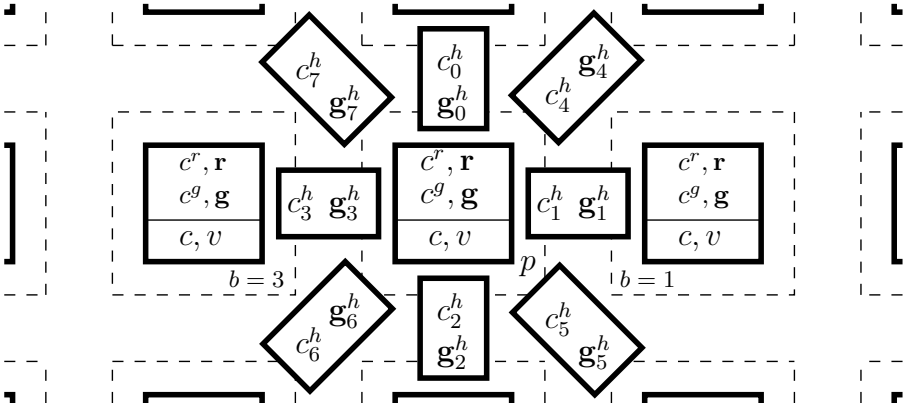


Abbildung 5.7: Repräsentationsschema für die $\mathcal{H}$ Diffusion. Für eine Zellposition p werden zunächst aus den Differenzen der Ausprägungen v zusammen mit den Konfidenzen c lokale Gradienten $\mathbf{g}_b^h$ und deren Konfidenz c_b^h auf den Zellgrenzen berechnet. Zusammen mit den Werten aus der Umgebung werden die Zellwerte $\mathbf{g}$, $\mathbf{r}$ und deren Konfidenzen c^g, c^r berechnet.

$\mathbf{g}^\delta$ anders aufgeteilt, so dass die auszutauschenden Konfidenzanteile auch von der Stärke des lokalen Gradienten abhängen. Dies wird auch auf allen anderen Schichten so durchgeführt.

Es ergibt sich eine neue c^h Konfidenz und ein neuer Gradient $\mathbf{g}^h$. Dieser hat die konstante Länge eins und behält nur die Richtung von g^δ in Bezug auf die Richtung der Nachbarschaftszelle, die mit $\mathbf{e}_b$ gegeben ist. Die Konfidenz c^h berücksichtigt dafür noch den Betrag der ursprünglichen Gradienten $\mathbf{g}_b^\delta$:

$$c_b^h(l,p) = c_b^\delta(l,p)\ ||\mathbf{g}_b^\delta(l,p)||, \tag{5.16}$$

$$\mathbf{g}_b^h(l,p) = \operatorname{sgn}\big(g_b^\delta(l,p)\big)\ \mathbf{e}_b. \tag{5.17}$$

Bezüglich der Diffusionsiterationen sind diese Werte statisch. Abbildung 5.7 zeigt, wie diese Werte im Raster angeordnet sind.

Um nun die für die Konfidenzdiffusion (5.13) verwendeten Anteile $\dot{c}^u$, $\hat{c}^u$, $\mathring{c}^u$ zu berechnen, werden an einem Zellübergang drei Werte ausgewertet: die lokale (kumulierende) Konfidenz c^g, die eingeleitete Konfidenz aus dem Zellübergang c^h und die Konfidenz aus der Umgebung (aus der Nachbarzelle) $c^g(p_b)$.

Würde man diese Werte direkt für $\dot{c}^u$, $\hat{c}^u$, $\mathring{c}^u$ einsetzen, wäre die Diffusion instabil. c^h würde konstant neue Anteile liefern, die sich in c^g in Abhängigkeit der Iterationsanzahl T^h beliebig kumulierten. Um das zu vermeiden, wird eine Stabilitätsbedingung aufgestellt. Sie drückt aus, dass die Summe

der verwendeten Lokal-, Rand- und Umgebungsanteile $\dot{c}_b^u$, $\hat{c}_b^u$, $\mathring{c}_b^u$ nie größer sein darf, als das Maximum der Anteile c^g, c_b^h und $c^g(p_b)$.

$$\dot{c}_b^u + \hat{c}_b^u + \mathring{c}_b^u \overset{!}{<} \max\left(c^g, c_b^h, c^g(p_b)\right). \tag{5.18}$$

Im Folgenden wird die Berechnung der verwendeten Anteile $\dot{c}^u$, $\hat{c}^u$, $\mathring{c}^u$ anhand von Relativierungen und Fallunterscheidungen aus c^g, c^h und $c^g(p_b)$ dargestellt. Letztere werden fortan als absolute Anteile bezeichnet.

Der lokale Anteil c^g dient bei den Simulationen als kumulierendes „Gedächtnis“. Wenn diese Historie korrekt berücksichtigt werden soll, muss der c^g Anteil für den nächsten Simulationsschritt unverändert zur Verfügung stehen. Es ergibt sich also eine Kontinuitätbedingung, die auch gleich den verwendeten Anteil der lokalen Gradientenkonfidenz $\dot{c}_b^u(l,p,t)$ liefert:

$$\dot{c}_b^u(l,p,t) \;=\; c^g(l,p,t), \tag{5.19}$$

Ein Fluss von Konfidenzanteilen in das Gedächtnis kommt nur zustande, wenn c^h oder $c^g(p_b)$ größer als c^g sind; bzw. die Differenzen $c^h - c^g$ oder $c^g(p_b) - c^g$ positiv sind. Die Größe der Differenzen soll die Flussmenge bestimmen.

Ist eine Konfidenz auf dem Zellübergang gegeben, soll sie stets Vorrang haben gegenüber einer Konfidenz aus der Umgebung. Würde diese Vorrangsbedingung nicht eingehalten, würde es noch nicht zwangsläufig zu einer Instabilität der Amplitude kommen - in der später beschriebenen Diffusion des Gradienten würden jedoch die ursprünglich eingeleiteten Anteile in der Umgebung verlaufen. Die Strukturen würden komplett vermischen, wenn T^h gegen unendlich geht. Die Vorrangsbedingung sorgt dafür, dass die Information von starken c^h auf den Zellgrenzen im Gedächtnis c^g erhalten bleiben.

Die folgenden Schritte ermöglichen die Berechnung von $\dot{c}_b^u$, $\hat{c}_b^u$ und $\mathring{c}_b^u$ unter Einhaltung der Stabilitäts- Kontinuitäts- und Vorrangsbedingung. Aus der Kontinuitätsbedingung geht hervor, dass c^g in voller Höhe erhalten bleibt. Die Stabilitätsbedingung schreibt dann vor, dass die Anteile c^h von der Zellgrenze und $c^g(p_b)$ aus der Umgebung nur noch einen Beitrag liefern können, wenn sie größer als c^g sind. Daher werden erstmal Differenzen dieser Anteile zu c^g gerechnet. Die Bezeichner der auf c^g relativierten Anteile werden mit einem Δ gekennzeichnet. Zur Einhaltung der Vorrangsbedingung wird der Anteil auf dem Zellübergang $c^{h\Delta}$ vorrangig berechnet:

$$c_b^{h\Delta}(l,p,t) \;=\; c_b^h(l,p) - c^g(l,p,t). \tag{5.20}$$

Der Anteil aus der Umgebung $c_g(p_b)$ kann nun nur noch einen Beitrag liefern, wenn er größer als c^h ist.

Zur Bestimmung von $\mathring{c}_b^u$ wird zunächst das Maximum $c^{\max}$ von Rand- und Umgebungsteil berechnet:

$$c_b^{\max}(l,p,t) \;=\; \max\big(c_b^h(l,p), c^g(l,p_b,t)\big). \tag{5.21}$$

Dies wird danach auf c^g relativiert:

$$c_b^{\max\Delta}(l,p,t) \;=\; c_b^{\max}(l,p,t) - c^g(l,p,t). \tag{5.22}$$

Danach wird der Umgebungsanteil $c^{g\Delta}$ aus dem „Rest“ berechnet, der sich aus dem Maximumsanteil abzüglich des Randanteils ermitteln lässt:

$$c_b^{g\Delta}(l,p,t) \;=\; c_b^{max\Delta}(l,p,t) - c_b^{h\Delta}(l,p,t). \tag{5.23}$$

Abschließend muss noch Sorge getragen werden, dass sich keine negativen Anteile bilden:

$$\hat{c}_b^u(l,p,t) \;=\; \begin{cases} c_b^{h\Delta}(l,p,t) & \text{für } \; c_b^{h\Delta}(l,p,t) > 0 \;\wedge\; c_b^{\max\Delta}(l,p,t) > 0 \\ 0 & \text{sonst} \end{cases} \tag{5.24}$$

$$\mathring{c}_b^u(l,p,t) \;=\; \begin{cases} c_b^{g\Delta}(l,p,t) & \text{für } \; c_b^{g\Delta}(l,p,t) > 0 \;\wedge\; c_b^{\max\Delta}(l,p,t) > 0 \\ 0 & \text{sonst} \end{cases} \tag{5.25}$$

Damit sind die in (5.13) verwendeten Anteile bestimmt. Abbildung 5.8 zeigt die Berechnung exemplarisch für drei wesentliche Fälle. In Abbildung 5.9 wird dargestellt, wie man sich die $\mathcal{H}$-Diffusion auch als Austausch von unterschiedlich gefärbten Flüssigkeiten vorstellen kann.

Die Berechnung des Gradienten $\mathbf{g}$ erfolgt nach Gleichung (5.12) anhand der Gradientenkomponenten $\dot{\mathbf{g}}(l,p,t)$, $\hat{\mathbf{g}}(l,p,t)$ und $\mathring{\mathbf{g}}(l,p,t)$. Deren Flussanteile werden anhand der zugehörigen Konfidenzen $\dot{c}^g(l,p,t)$, $\hat{c}^g(l,p,t)$ und $\mathring{c}^g(l,p,t)$ vorgenommen. Diese werden noch mit der Konfidenz im Folgeschritt $c^g(l,p,t+1)$ (also der Summe der Konfidenzanteile) normiert:

$$\begin{aligned} \mathbf{g}(l,p,t+1) \;&=\; \dot{\mathbf{g}}(l,p,t) + \hat{\mathbf{g}}(l,p,t) + \mathring{\mathbf{g}}(t) \;=\; \\ &\underbrace{\sum_b \sigma_b \frac{\dot{c}_b^g(l,p,t)}{c^g(l,p,t+1)}\,\mathbf{g}(l,p,t)}_{\dot{\mathbf{g}}(l,p,t)} + \\ &\underbrace{\sum_b \sigma_b \frac{\hat{c}_b^g(l,p,t)}{c^g(l,p,t+1)}\,\mathbf{g}_b^h(l,p,t)}_{\hat{\mathbf{g}}(l,p,t)} + \\ &\underbrace{\sum_b \sigma_b \frac{\mathring{c}_b^g(l,p,t)}{c^g(l,p,t+1)}\,\mathbf{g}(l,p_b,t)}_{\mathring{\mathbf{g}}(l,p,t)} \end{aligned} \tag{5.26}$$

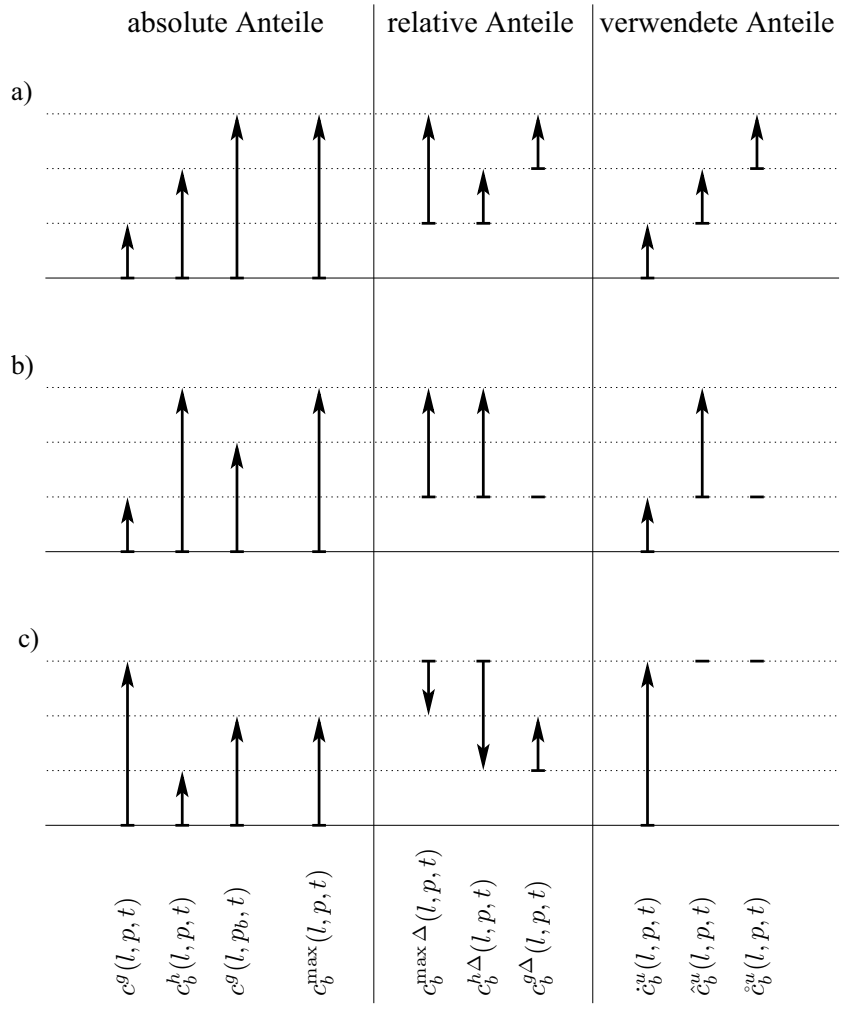


Abbildung 5.8: Berechnung der Konfidenzanteile der $\mathcal{H}$-Diffusion. Die Berechnung der Anteile $\dot{c}^u$, $\hat{c}^u$, $\mathring{c}^u$ die zur Ermittlung der Gradientenkonfidenz Verwendung finden, erfolgt unter Einhaltung einer Kontinuitäts-, einer Stabilitäts- und einer Vorrangsbedingung (siehe Text). Die Berechnung wird für drei Konstellationen von Werten gezeigt. Betrachtet wird nur ein Übergang zu einer Nachbarzelle in der Richtung b. Im Fall a) ist der lokale (kumulierende) Anteil c^g gering. Der Anteil von der Zellgrenze c_b^h und der Umgebung $c^g(p_b)$ sind größer. Die relativen Anteile werden auf c^g bezogen. Dabei erreicht $c_b^{h\Delta}$ die Höhe von c_b^h. Der Rest bleibt für den Anteil aus der Umgebung $c_b^{g\Delta}$. Die verwendeten Anteile sind entsprechend. Im Fall b) ist der Anteil auf dem Zellübergang c_b^h größer als der aus der Umgebung $c^g(p_b)$. Ersterer bestimmt dann auch den verwendeten Anteil. D.h. in diesem Fall hat die auf dem Zellübergang eingeleitete Information vorrang. Das ist wichtig, weil sonst die eingeleitete Information im Laufe der Iterationen nicht an der ursprünglichen Position bleiben würde. Im Fall c) ist die lokale Konfidenz c^g größer als die anderen absoluten Anteile. Es wird dann auch nur dieser Anteil verwendet.

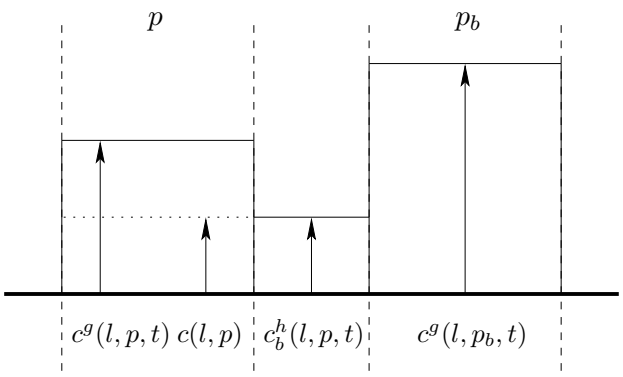


Abbildung 5.9: Die $\mathcal{H}$-Diffusion kann auch als Austausch von Flüssigkeiten interpretiert werden. Zwischen den Behältern der Zellen befinden sich noch Behälter, deren Füllhöhe den Randwerten entspricht. Sie leiten sich aus Differenzen der Ausprägungen und deren Konfidenzen ab und stellen in diesem Modell Quellen dar, die stets ihr Niveau konstant halten. Die auszutauschenden Informationen werden aus den Differenzen der Füllhöhen bestimmt. Dabei haben die Randwerte c^h Vorrang gegenüber den Werten $c^g(p_b)$ aus der Umgebung.
Daraus berechnet sich dann eine Gradientenkonfidenz im Folgeschritt: $c^g(l,p,t+1)$. Sie kumuliert alle bisherigen ausgetauschen Anteile. Der Gradient kann als Farbe (Richtung) und Sättigung (Länge) gedeutet werden, der entsprechend der ausgetauschten Konfidenzanteile gemischt wird.

$$\begin{aligned}\mathbf{g}(l,p,t+1) \;=\; & \\ & \frac{1}{c^g(l,p,t+1)}\Big(c^g(l,p,t)\,\mathbf{g}(l,p,t) + \\ & \qquad \sum_b \sigma_b\big(\hat{c}_b^g(l,p,t)\,\mathbf{g}_b^h(l,p,t) + \mathring{c}_b^g(l,p,t)\,\mathbf{g}(l,p_b,t)\big)\Big).\end{aligned} \tag{5.27}$$

Für die Berechnung der Translation wird für alle Richtungen b ausgewertet, welche Anteile zum neuen Gradient $\mathbf{g}(l,p,t+1)$ beigetragen haben. Die Gewichtungsfaktoren sind daher identisch mit denen aus der Berechnung des Gradienten. Der $\dot{\mathbf{r}}$ Anteil wirkt wieder kumulierend, während der $\mathring{\mathbf{r}}$ Anteil die Translationsanteile aus der Umgebung aufnimmt und um den zugehörigen Rasterabstand mit $\sigma_b\mathbf{e}_b$ erhöht. Der $\hat{\mathbf{r}}$ Anteil ist null. Lokale Gradientenanteile wirken als Nullanteil, da sie keine Entfernung zur auswertenden Zelle haben. Durch die Normierung der Gradientenkonfidenz über alle Anteile, werden bei Gradienten auf den Zellgrenzen dann aber die anderen Translationskomponenten geschwächt. Es ergibt sich:

$$\mathbf{r}(l,p,t+1) \;=\; \dot{\mathbf{r}}(l,p,t) + \hat{\mathbf{r}}(l,p,t) + \mathring{\mathbf{r}}(l,p,t) \;=$$

$$\underbrace{\sum_b \sigma_b \frac{\dot{c}_b^g(l,p,t)}{c^g(l,p,t+1)}\,\mathbf{r}(l,p,t)}_{\dot{\mathbf{r}}(l,p,t)} + \underbrace{\mathbf{0}}_{\hat{\mathbf{r}}(l,p,t)} + \tag{5.28}$$

$$\underbrace{\sum_b \sigma_b \frac{\mathring{c}_b^g(l,p,t)}{c^g(l,p,t+1)}\,\big(\mathbf{r}(l,p_b,t)+\mathbf{e}_b\big)}_{\mathring{\mathbf{r}}(l,p,t)}$$

$$= \;\frac{1}{c^g(l,p,t+1)}\sum_b \sigma_b\Big(\dot{c}_b^g(l,p,t)\,\mathbf{r}(l,p,t) \;+ \mathring{c}_b^g(l,p,t)\,\big(\mathbf{r}(l,p_b,t)+\mathbf{e}_b\big)\Big)$$

Auch die Translationskonfidenz c^r berechnet sich aus einem kumulierenden Anteil $\dot{c}^r$ und einem Anteil aus der Umgebung $\mathring{c}^r$. Die Gewichtung erfolgt ebenfalls über die normierte Gradientenkonfidenz und vernächlässigt den Anteil aus der Zellengrenze. Der Anteil aus der Umgebung vergleicht, wie groß die Übereinstimmung des hinzugeflossenen Gradienten mit dem bisher kumulierten Gradienten ist. Dies erfolgt mit dem normierten Skalarprodukt. Ist die Übereinstimmung groß, wird davon ausgegangen, dass die zugehörigen Translationswerte auch den gleichen Ursprung haben und die Konfidenz erhöht sich entsprechend:

$$c^r(l,p,t+1) \;=\; \dot{c}^r(l,p,t) + \hat{c}^r(l,p,t) + \mathring{c}^r(l,p,t) \;=$$

$$\underbrace{\frac{\dot{c}^g(l,p,t)}{c^g(l,p,t+1)}c^r(l,p,t)}_{\dot{c}^r(l,p,t)} + \underbrace{\mathbf{0}}_{\hat{c}^r(l,p,t)} + \tag{5.29}$$

$$\underbrace{\sum_b \sigma_b \frac{\mathring{c}^g(l,p,t)}{c^g(l,p,t+1)}\,\frac{\langle \dot{\mathbf{g}}(l,p,t), \mathring{\mathbf{g}}_\mathbf{b}(l,p,t)\rangle}{||\dot{\mathbf{g}}(l,p,t)||\,||\mathring{\mathbf{g}}_\mathbf{b}(l,p,t)||}}_{\mathring{c}^r(l,p,t)}$$

$$= \;\frac{1}{c^g(l,p,t+1)}\Big(\dot{c}^g(l,p,t)\,c^r(l,p,t) + \tag{5.30}$$

$$\sum_b \sigma_b \mathring{c}^g(l,p,t)\,\frac{\langle \dot{\mathbf{g}}(l,p,t), \mathring{\mathbf{g}}_\mathbf{b}(l,p,t)\rangle}{||\dot{\mathbf{g}}(l,p,t)||\,||\mathring{\mathbf{g}}_\mathbf{b}(l,p,t)||}\Big)$$

Hat der Iterationsindex t den Maximalwert T^h erreicht, dann werden

die Iterationswerte für die folgenden Verarbeitungsschritte übernommen:

$$c^g(l,p) = c^g(l,p,t=T^h), \quad (5.31)$$
$$\mathbf{g}(l,p) = \mathbf{g}(l,p,t=T^h), \quad (5.32)$$
$$c^r(l,p) = c^r(l,p,t=T^h), \quad (5.33)$$
$$\mathbf{r}(l,p) = \mathbf{r}(l,p,t=T^h). \quad (5.34)$$

Damit sich stabile Repräsentationen der Iterationswerte einstellen können, wird das Diffusionsverfahren für mehrere Iterationsschritte nacheinander durchgeführt.

Wie beim Diffusionsverfahren zur Interpolation hängt eine sinnvolle Anzahl von Iterationsschritten von der Bildgröße und vom Bildinhalt ab. Auch hier ist ein sicherer Wert die Anzahl der Pixel in der größten Bilddimension.

Wie das Diffusionsverfahren zur Interpolation ist auch dieses parameterfrei. Eine implizite Parametrisierung ergibt sich hier aus der Gleichgewichtung der unterschiedlichen lokalen Anteile, der Anteile aus den Zellgrenzen und Anteile aus der Umgebung.

5.1.4 Zustandsvektor

Für jeden Rasterpunkt an einer Position p einer Schicht l wird ein (Teil-)-Zustandsvektor $\mathbf{x}^l$ gebildet. Dieser enthält die lokale Ausprägung v und die Vektorkomponenten der Strukturbeschreibungen $\mathbf{r}$ und $\mathbf{g}$:

$$\mathbf{x}^l(l,p) = \big(v\big(l,p\big), r_1(l,p), r_2(l,p), g_1(l,p), g_2(l,p)\big)^T. \quad (5.35)$$

Die Indexe 1 und 2 beschreiben kartesische Vektorkomponenten in den Bilddimensionen. Die Translationskomponenten wachsen mit dem Abstand des dominierenden Gradienten. Dadurch können sich in Abhängigkeit vom Hintergrund der zu lernenden Region sehr unterschiedliche Werte in Hintergrundbereichen ergeben. Es wurden $1/x$ und $Exp(-x)$ Repräsentationen der Komponenten überprüft. Es wurde auch versucht, die Transformationen erst nach der Rücktransformation noch in Polarrepräsentation nur auf die Länge anzuwenden. Die besten Ergebnisse ergaben sich jedoch in der in Gleichung (5.35) dargestellten Weise, wenn die Lernregion auf das Objekt begrenzt wurde.

Die zu den Vektorkomponenten zugehörigen Konfidenzen werden mit einem Vektor $\mathbf{c}^l$ beschrieben:

$$\mathbf{c}^l(l,p) = \big(c(l,p), c^r(l,p), c^r(l,p), c^g(l,p), c^g(l,p)\big)^T. \quad (5.36)$$

Die strukturelle Konfidenz c^g wird für **r**- und **g**-Komponenten jeweils identisch eingetragen. So ergibt sich $\mathbf{c}^l$ mit korrespondierenden Komponenten zu $\mathbf{x}^l$, was im Folgenden Ausdrücke vereinfacht.

Für die Assoziation wird ein Zustandsvektor $\mathbf{x}^p$ zusammengesetzt, der alle (Teil-)Zustandsvektoren $\mathbf{x}^l$ unterliegender Schichten enthält:

$$\mathbf{x}^p(l,p) = \begin{cases} \mathbf{x}^l\Big(\lambda\big(h(l)-1\big),p\Big) \circ \mathbf{x}^p\Big(\lambda\big(h(l)-1\big),p\Big) & \text{für} \quad h(l) \geq 1 \\ () & \text{für} \quad h(l) = 0 \end{cases} \tag{5.37}$$

$$\text{mit} \;\; \mathbf{x}^p\Big(\lambda\big(0\big),p\Big) = () \;\; \text{und der Vektorkonkatenation}$$

$$\mathbf{x} \circ \mathbf{y} = (x_1, x_2, \cdots, x_{|\mathbf{x}|}, y_1, y_2, \cdots, y_{|\mathbf{y}|}). \tag{5.38}$$

Entsprechend ergibt sich ein Konfidenzvektor $\mathbf{c}^p$:

$$\mathbf{c}^p(l,p) = \begin{cases} \mathbf{c}^l\Big(\lambda\big(h(l)-1\big),p\Big) \circ \mathbf{c}^p\Big(\lambda\big(h(l)-1\big),p\Big) & \text{für} \quad h(l) \geq 1 \\ () & \text{für} \quad h(l) = 0 \end{cases} \tag{5.39}$$

$$\text{mit} \;\; \mathbf{c}^p\Big(\lambda\big(0\big),p\Big) = ()$$

Die Komponenten der Vektoren $\mathbf{x}^p$ und $\mathbf{c}^p$ werden mit m indiziert. Die Anzahl M von Komponenten ist abhängig von der Schicht:

$$M(l) = \dim\big(\mathbf{x}^p(l,.)\big) = \dim\big(\mathbf{x}^l(l,.)\big)\big(h(l)-1\big) = 5\big(h(l)-1\big). \tag{5.40}$$

Das Provadero-Verfahren verwendet Schichtengruppen. Jeder Schicht einer Schichtengruppe ist ein Konzept zugeordnet. Damit alle Konzepte auf den gleichen Zustandsvektorbereich zurückgreifen können, wird ein Zustandsvektor $\mathbf{x}^\Omega$ definiert, der alle Schichten(gruppen) unterhalb einer Schichtengruppe Ω beinhaltet.

$$\mathbf{x}^\Omega(l,p) = \mathbf{x}^p(\lambda(\Omega),p) \tag{5.41}$$

Die Funktion $\lambda(\Omega)$ liefert die unterste Schicht der Schichtengruppe Ω

$$\lambda(\Omega) = l \in \Omega \mid \lambda(l) \leq \lambda(l') \;\; \forall\, l' \in \Omega. \tag{5.42}$$

Der Zustandskonfidenzwektor $\mathbf{c}^\Omega$ berechnet sich analog zu (5.41):

$$\mathbf{c}^\Omega(l,p) = \mathbf{c}^p(\lambda(\Omega),p) \tag{5.43}$$

Die Anzahl M der Konponenten bestimmt sich aus:

$$\begin{aligned} M(\Omega) &= \dim\big(\mathbf{x}^\Omega(l)\big) \;\text{mit} \quad l \in \Omega \\ &= 5\,\Big(\big(h(\Omega)-2\big)\,|O| + \#\text{Eingabebildkanäle}\Big). \end{aligned} \tag{5.44}$$

5.1.5 Freigabe

Die Freigabefunktion $\mathcal{E}$ wird genutzt, um nur die Positionen der Vektoren $\mathbf{x}^{\Omega}$ weiterzuleiten, die für die auswertende Schicht l von Bedeutung sind. Realisiert ist dies über die Multiplikation mit einer Matrix E, die für jede Schicht l separat bestimmt wird. Sie berechnet einen neuen Zustandsvektor $\mathbf{x}^e$ und eine zugehörige Konfidenz. Beide haben die Dimensionalität $N(l)$, die innerhalb einer Schichtengruppe einheitlich ist:

$$
\begin{aligned}
N(l) &= N(\Omega)\ \forall\, l \in \Omega && (5.45)\\
\mathcal{E}(l) &: \mathbb{R}^{M(\Omega)} \rightarrow \mathbb{R}^{N(l)} \ \big|\ l \in \Omega; && (5.46)\\
\mathbf{x}^e(l,p) &= E(l)\,\mathbf{x}^{\Omega}(l,p) && (5.47)\\
\mathbf{c}^e(l,p) &= E(l)\,\mathbf{c}^{\Omega}(l,p). && (5.48)
\end{aligned}
$$

Die Matrix enthält binäre Elemente und ist im nicht reduzierenden Fall eine Identitätsmatrix. Für jede zu löschende Dimension wird einfach die entsprechende Zeile gelöscht. Es ergibt sich eine entsprechend reduzierte Anzahl N von Komponenten. In Abschnitt 5.2.1 wird beschrieben, wie $E(l)$ so eingestellt wird, dass nur Komponenten zur Auswertung gelangen, die bezüglich der gelernten Konzepte möglichst diskriminativ sind. Für Komponenten, die aus (kartesischen) Vektorbeschreibungen hervorgegangen sind, werden die Vektorpositionen stets paarweise freigegeben.

Vektorpositionen von Zustandsvektoren vor der Freigabe werden im Folgenden mit m indiziert. Vektorpositionen von Zustandsvektoren nach der Freigabe werden mit $n$indiziert.

5.1.6 Rücktransformation

Die Rücktransformation $\mathcal{B}$ soll den Zustandsvektor $\mathbf{x}^e$ der assoziierenden Schicht so verändern, dass die Assoziation möglichst invariant erfolgen kann. Als Ergebnis wird der Zustandsvektor $\mathbf{x}^b$ mit zugehöriger Konfidenz $\mathbf{c}^b$ berechnet. Die Rücktransformation wird parametrisiert durch die interpolierten Ausprägungen v^f unterliegender Schichten und den zugehörigen Konfidenzen c^f. Diese interpolieren die v-Werte und wurden eingeführt, da v-Werte typischerweise nur spärlich aktiviert sind. Mit den v-Werten könnten nur Inselbereiche rücktransformiert werden. Mit den v^f-Werten ist dagegen eine flächendeckende Rücktransformation möglich. Das ist z.B. für einen rotierten Flächenbereich sinnvoll, der keine innere Bildstruktur zeigt, aber trotzdem als ganzer Flächenbereich rücktransformiert werden soll.

$$\mathcal{B}(\Omega) \quad : \quad \mathbb{R}^{N(\Omega)} \times \mathbb{R}^{M(\Omega)} \times \mathbb{R}^{M(\Omega)} \rightarrow \mathbb{R}^{N(\Omega)}; \tag{5.49}$$

$$\mathbf{x}^b(l,p) \quad = \quad \mathcal{B}\Big(\mathbf{x}^e(l,p), \big\{v^f(l',p)\big\}, \big\{c^f(l',p)\big\} \mid h(l') < h(\Omega)\Big) \tag{5.50}$$

$$\mathbf{c}^b(l,p) \quad = \quad \mathbf{c}^e(l,p) \tag{5.51}$$

Bevor die eigentliche Rücktransformation erfolgt, werden die im Zustandsvektor $\mathbf{x}^e$ enthaltenen Signale, die Vektorkomponenten kodieren in Polarkoordinaten gewandelt. Das ist sinnvoll, weil Bildtransformationen in diesen Koordinaten häufig gut zu separieren sind und weil die Transformationseigenschaften so anschaulicher sind. Die Wandlung φ erzeugt einen neuen Zustandsvektor $\mathbf{x}^\varphi$:

$$\mathbf{x}^\varphi(l,p) \quad = \quad \varphi\big(\mathbf{x}^e(l,p)\big) \tag{5.52}$$

Für Komponenten, die Ausprägungen kodieren, ergibt sich keine Änderung:

$$x_n^\varphi(l,p) \quad = \quad x_n^e(l,p) \tag{5.53}$$

Komponenten, die Vektoren (Gradient oder Translation) kodieren, werden in Polarkoordinaten gewandelt:

$$x_n^\varphi(l,p) \quad = \quad \sqrt{\big(x_n^e(l,p)\big)^2 + \big(x_{n+1}^e(l,p)\big)^2} \tag{5.54}$$

$$x_{n+1}^\varphi(l,p) \quad = \quad \mathrm{arctan2}\big(x_n^e(l,p), x_{n+1}^e(l,p)\big). \tag{5.55}$$

Dabei wird die häufig zur Polarkoordinatenberechnung Funktion arctan2 verwendet:

$$\mathrm{arctan2}(i,j) \quad = \quad \begin{cases} \arctan\frac{j}{i} & \text{für } i > 0 \\ \arctan\frac{j}{i} + \pi & \text{für } i < 0, j \geq 0 \\ \arctan\frac{j}{i} - \pi & \text{für } i < 0, j < 0 \\ +\frac{\pi}{2} & \text{für } i = 0, j > 0 \\ -\frac{\pi}{2} & \text{für } i = 0, j < 0 \\ 0 & \text{für } i = 0, j = 0 \end{cases} \tag{5.56}$$

Die eigentliche Transformation erfolgt linear mit einem additiven Anteil β^o und einem multiplikativen Anteil β^s. Die Funktion ist so ausgelegt, das keine Rücktransformation erfolgt, wenn diese Werte null sind:

$$x_n^\beta(l,p) \quad = \tag{5.57}$$

$$\begin{cases} \Big(\beta_n^o(l,p) + \big(1 + \beta_n^s(l,p)\big)\, x_n^\varphi(l,p)\Big) \bmod 2\pi & : \text{für Winkelkomponenten} \\ \Big(\beta_n^o(l,p) + \big(1 + \beta_n^s(l,p)\big)\, x_n^\varphi(l,p)\Big) & : \text{für alle anderen Komp.} \end{cases}$$

Die Transformationsparameter β^o und β^s sind von den Ausprägungen v^f und Konfidenzen c^f unterliegender Schichten abhängig:

$$\beta_n^o(l,p) = \frac{1}{\sum_{l'|h(l')<h(\Omega)} c^f(l',p)} \sum_{l''|h(l'')<h(\Omega)} c^f(l'',p)\; v^f(l'',p)\; b_n^o(l,l''), \tag{5.58}$$

$$\beta_n^s(l,p) = \frac{1}{\sum_{l'|h(l')<h(\Omega)} c^f(l',p)} \sum_{l''|h(l'')<h(\Omega)} c^f(l'',p)\; v^f(l'',p)\; b_n^s(l,l''). \tag{5.59}$$

Die Faktoren b^o und b^s bestimmen die Rücktransformationseigenschaften einer Schicht. Sie sind interne Parameter der Rücktransformation und werden mit einem in Abschnitt 5.2.2 beschriebenen Lernverfahren für jede Schicht eingestellt.

Abschließend werden die Koordinaten wieder in kartesiche Koordinaten mit der Funktion φ^{-1} zurückgewandelt:

$$\mathbf{x}^b(l,p) = \varphi^{-1}\big(\mathbf{x}^\beta\big) \tag{5.60}$$

Dabei bleiben die Ausprägungskomponenten wieder unverändert:

$$x_n^b(l,p) = x^\beta(l,p) \tag{5.61}$$

Vektorkomponenten werden wie folgt bestimmt:

$$x_n^b(l,p) = x_n^\beta(l,p)\; \cos\big(x_{n+1}^\beta(l,p)\big) \tag{5.62}$$

$$x_{n+1}^b(l,p) = x_n^\beta(l,p)\; \sin\big(x_{n+1}^\beta(l,p)\big). \tag{5.63}$$

5.1.7 Skalierung

Danach wird der Zustandsvektor mit der Funktion $\mathcal{S}$ skaliert. Das ist notwendig, da die Struktur beschreibenden Variablen unterschiedliche Wertebereiche haben und für die anschließende Projektion mittelwertfreie Variablen mit identischer Varianz benötigt werden.

Die im nächsten Abschnitt beschriebene Projektion vermisst Abstände in einem Merkmalsraum, der durch die Komponenten des Zustandsvektors aufgespannt wird. Damit die Komponenten dabei im Mittel gleiche Beiträge liefern können, werden sie auf ihr Moment zweiter Ordnung normiert. Dazu müssen sie vorab noch mittelwertfrei gemacht werden. Die Skalierung $\mathcal{S}$ erzeugt einen neuen Zustandsvektor $\mathbf{x}^s$ mit zugehöriger Konfidenz $\mathbf{c}^s$:

$$\mathcal{S}(\Omega) : \mathbb{R}^{N(\Omega)} \to \mathbb{R}^{N(\Omega)}; \tag{5.64}$$

$$\mathbf{x}^s(l,p) = \mathcal{S}\big(\mathbf{x}^b(l,p)\big) \tag{5.65}$$

$$\mathbf{c}^s(l,p) = \mathbf{c}^b(l,p) \tag{5.66}$$

Die Komponenten werden wie folgt berechnet:

$$x_n^s(l,p) \;=\; \frac{x_n^b(l,p) - m_n^1(l)}{\sqrt{m_n^2(l)}} \tag{5.67}$$

Die Momentvektoren $\mathbf{m}^1$ und $\mathbf{m}^2$ werden mit einem Lernverfahren eingestellt, dass in Abschnitt 5.2.3 beschrieben wird.

5.1.8 Projektion

Die Projektion $\mathcal{P}$ beschreibt, wie an einer Rasterposition p einer Schicht l aus dem Zustandsvektor $\mathbf{x}^s(l,p)$ die Werte der Ausprägung v und der Konfidenz c berechnet werden.

$$\mathcal{P}(\Omega) \;:\; \mathbb{R}^{N(\Omega)} \times \mathbb{R}^{N(\Omega)} \rightarrow \mathbb{R}^2; \tag{5.68}$$

$$(c,v) \;=\; \mathcal{P}\big(\mathbf{x}^s(l,p), \mathbf{c}^s(l,p)\big) \tag{5.69}$$

Die Projektion verwendet als internen Parameter die Spezifität $\mathbf{s}(l)$. Sie beschreibt in welcher Richtung die Zustandsvektoren aus den Trainingsmustern die größte Varianz zeigen. Sie stellt damit die erste Haupkomponente[1] einer Haupkomponentenanalyse dieser Zustandsvektoren dar. Abschnitt 5.2.4 zeigt, wie diese berechnet wird.

Wird eine Abbildung assoziiert, wird die Länge v der Projektion von $\mathbf{x}^s$ auf $\mathbf{s}$ gemessen. Sie entspricht der Ausprägung der neuen Abbildung relativ zu den trainierten Mustern. Die Aktivierung $\tilde{c}$ soll hohe Werte zugewiesen bekommen, wenn der Abstand $||\mathbf{d}||$ klein ist und entsprechend umgekehrt. Der Abstand $||\mathbf{d}||$ beschreibt die Entfernung zwischen dem Zustandsvektor $\mathbf{x}^s$ und seinem Fußpunkt auf der Geraden, die durch die Spezifität $\mathbf{s}$ verläuft. Ist der Abstand klein, kann davon ausgegangen werden, dass der aktuelle Zustandsvektor gut zu den trainierten Vektoren passt. Abbildung 4.3 zeigt eine Beispielprojektion und Abbildung 5.10 zeigt, wie sich das Provadero-Detektionselement von einem klassischen unterscheidet. Die Berechnung von v erfolgt anhand der Projektion:

$$\frac{\mathbf{x}^s(l,p)^T \mathbf{s}(l)}{||\mathbf{x}^s(l,p)||\;||\mathbf{s}(l)||} \;=\; \cos\alpha(l,p) \;=\; \frac{v(l,p)}{||\mathbf{x}^s(l,p)||}. \tag{5.70}$$

Da $\mathbf{s}$ auf die Länge 1 normiert ist, ergibt sich:

$$v(l,p) = \mathbf{x}^s(l,p)^T \mathbf{s}(l). \tag{5.71}$$

[1] Die weiteren Hauptkomponenten scheinen für biologisch motivierte Informationsverarbeitung keine große Rolle zu spielen (Hyvärinen et al., 2009, S.99)

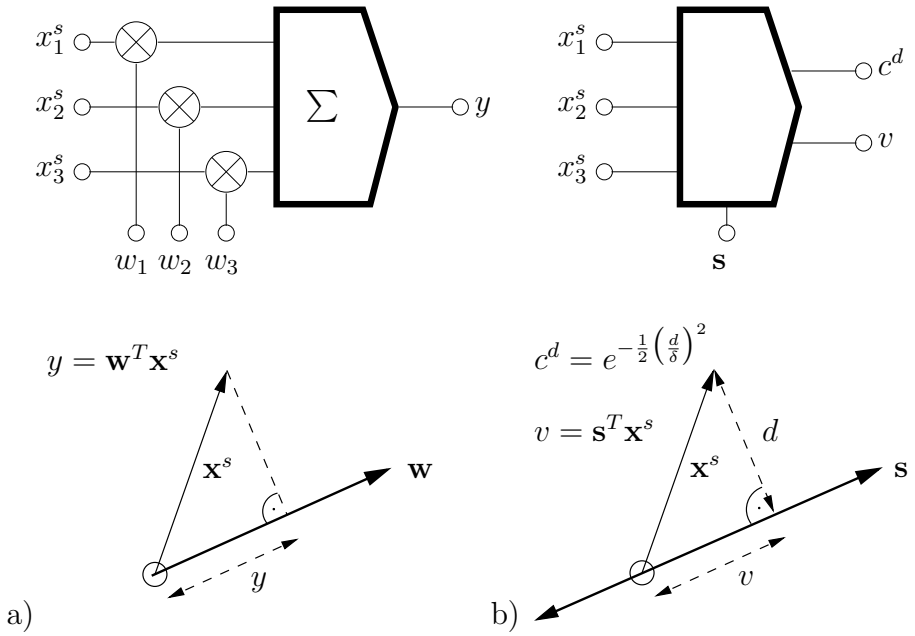


Abbildung 5.10: Klassisches- und Provadero-Detektionselement. Die Vektoren **w** und **s** sind Einheitsvektoren. Trotz gleicher Formeln für die Berechnung von y und v ergeben sich unterschiedliche Assoziationsverhalten aus unterschiedlichem Lernen für **w** und **s**. Das Provadero-Detektionselement berechnet die Ausprägung v und deren Konfidenz c anhand einer Projektion des Zustandsvektors $\mathbf{x}^s$ auf die Spezifität **s**. Für große Werte von d ergibt sich eine geringe Konfidenz c.

Für die Konfidenz wird zunächst ein Zwischenwert $\tilde{c}$ berechnet, der dann im nächsten Verarbeitungsschritt zur Konfidenz c normiert wird. Die Konfidenz $\tilde{c}$ hängt von dem Projektionsabstand **d** ab. Dieser berechnet sich aus:

$$\mathbf{d}(l,p) \;=\; \mathbf{x}^s(l,p) - v(l,p)\,\mathbf{s}(l) \tag{5.72}$$

Eine Projektion, die gut an den bisher trainierten Mustern und damit an **s** liegt, hat einen geringen Abstand **d**. Für eine solche Projektion soll sich eine hohe Konfidenz ergeben. Diese Abhängigkeit wird als negativ proportionale Funktion modelliert. Es ergibt sich die Konfidenz c^d:

$$c_n^d(l,p) \;=\; e^{-\frac{1}{2}\left(\frac{d_n}{\delta}\right)^2} \tag{5.73}$$

Der Faktor δ bestimmt den Radius von d_n, ab dem die Konfidenz zu null werden soll. Die Funktion ist beispielhaft in Abbildung 5.11 dargestellt. Zur

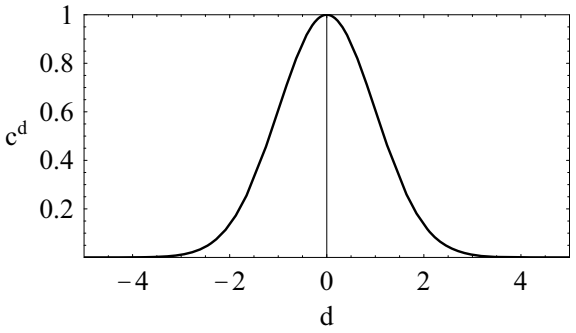


Abbildung 5.11:
Die Konfidenz c^d ist abhängig von dem Projektionsabstand d. Es wurde eine (symmerische) Gaußfunktion gewählt (hier beispielhaft mit $\delta = 1$ dargestellt).

Berechnung von $\tilde{c}$ erfolgt nun noch eine Gewichtung mit dem Konfidenzvektor $\mathbf{c}^s$. Das ist notwendig, um Werte zu eliminieren, die nur eine geringe Konfidenz haben und daher Ausreißer darstellen können. Schließlich erfolgt noch eine geometrische Mittelung:

$$\tilde{c}(l,p) = \frac{1}{\left|\left|\mathbf{c}^s(l,p)\right|\right|}\left|\left|\mathbf{c}^s(l,p) \odot \mathbf{c}^d\right|\right| \tag{5.74}$$

mit dem Vektorkomponentenprodukt

$$\mathbf{x} \odot \mathbf{y} = (x_1 y_1, x_2 y_2, \cdots, x_{|\mathbf{x}|} y_{|\mathbf{y}|}). \tag{5.75}$$

5.1.9 Einleitung eines Eingabebildes

Auf der untersten Schichtengruppe erfolgen die vorab dargestellten Assoziations-Verarbeitungsschritte nicht. Dort werden lediglich ein oder mehrere Kanäle des Eingabebildes angelegt. Die nächste Schichtengruppe arbeitet dann direkt auf diesen Daten. Eine Funktion $\mathcal{V}$ transformiert Pixelwerte eines Eingabebildes auf die Ausprägung v. Sie werden auf das Intervall $[-1/2..1/2]$ abgebildet. Die Konfidenz $\tilde{c}$ dieser Werte wird auf den Wert einer zum Bild gehörenden Maske gesetzt. Eine solche Maske enthielte den Wert "1" für Bereiche, auf dem ein zu lernendes Konzept sicher gegeben ist und den Wert "0" sonst. Auch für komplett assozierende Bilder kann die Maske verwendet werden, wenn z.B. ein anderer Umriß, als ein quadratischer gewünscht ist. Standardmäßig ist die Bildmaske homogen auf den Wert "1" gesetzt.

$$\begin{aligned}
\mathcal{V}(0) &: \mathbb{R} \rightarrow \mathbb{R}; && (5.76)\\
&: Pixelwert(p) \rightarrow v(l,p) \in \mathbb{R}, [-1/2..1/2] && (5.77)\\
\tilde{c}(l,p) &= Bildmaske(p) \in \{0,1\} && (5.78)
\end{aligned}$$

Für Farbbilder wird zunächst eine LAB Farbraumtransformation durchgeführt. Die Helligkeitskomponente L wird wie oben beschrieben auf der ersten Schicht dargestellt. Die Farbkomponenten A und B werden dann analog auf zwei zusätzlichen Schichten dargestellt. Die unterste Schichtengruppe enthält dann drei Schichten. Gegenüber anderen typischen Farbräumen stellt der LAB-Farbraum mit seiner Metrik die menschliche Wahrnehmung nach. Wenn z.B. zwei Punkte in diesem Raum als Farbmuster gegeben sind, dann entspricht die Länge des Abstands in diesem Raum der Intensität des wahrgenommenen Unterschieds. Eine solche Metrik ist generell für Verfahren zum Bildverstehen als Ausgangsbasis günstig, da die Erkennung von Bildstrukturen auch nach Helligkeits- und Farbunterschieden erfolgt, die dieser Metrik unterliegen.

5.1.10 Normierung

Die Funktion $\mathcal{N}(\Omega)$ normiert die Konfidenz $\tilde{c}$ innerhalb einer Schichtengruppe Ω. Die resultierende Konfidenz c beträgt eins, wenn über alle Muster q', alle Schichten l' und alle Pixelpositionen p' der Schichtengruppe gemittelt wird. Dieser Schritt ist wichtig, wenn Schicht für Schicht assoziiert wird und die Konfidenzen dabei stabil bleiben sollen. Wesentlich zum Konzepterkennen sind die relativen Unterschiede der Konfidenzen. Diese bleiben bei dieser Operation erhalten.

$$\mathcal{N}(\Omega) \quad : \quad \mathbb{R} \rightarrow \mathbb{R}; \tag{5.79}$$

$$c(q,l,p) \quad = \quad \tilde{c}(q,l,p)\, \frac{\sum_{q' \in Q} \sum_{l' \in \Omega} \sum_{p'} 1}{\sum_{q' \in Q} \sum_{l' \in \Omega} \sum_{p'} \tilde{c}(q',l',p')} \;\Bigg|\; l \in \Omega. \tag{5.80}$$

5.2 Lernen

Das Provadero-Verfahren wird anhand überwachter Lernverfahren für das Erkennen von einer Menge O von Konzepten o trainiert. Dazu werden eine Menge Q^L von Trainingsmustern q^L verwendet. Einem Muster sind ein Bild ψ und ein Konzept o zugeordnet: $q = \{\psi, o\}$. Für jedes Konzept sind verschiedene Trainingsmuster gegeben, die das Konzept in anwendungstypischen Varianzen darstellen. Nach dem Lernvorgang wird das Verfahren anhand einer Menge Q^T von Testmustern q^T evaluiert.

Der Ablauf des Lernverfahrens ist in Abbildung 5.12 dargestellt. Zunächst wird für alle Trainings- und Testmuster die unterste Schichtengruppe aus

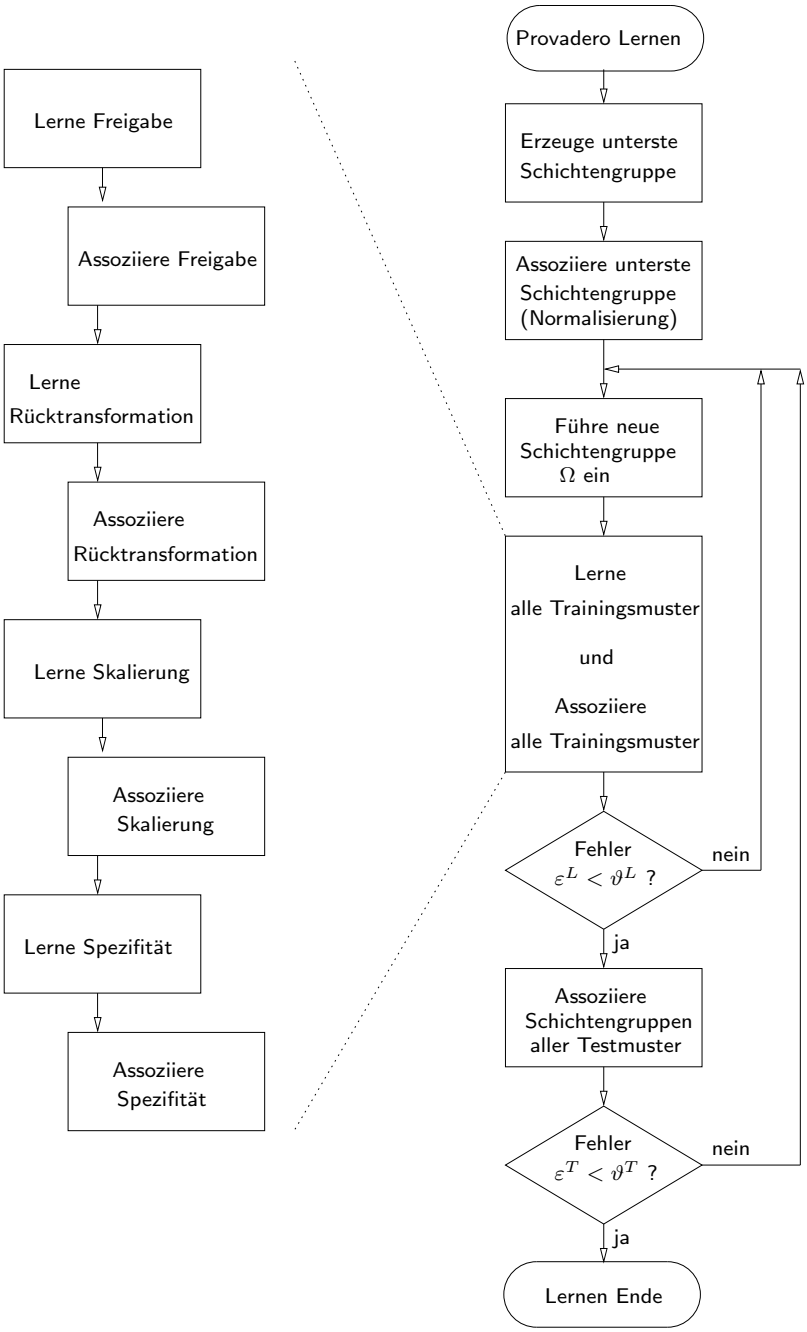


Abbildung 5.12: Flussdiagramm zum Provadero-Lernen (Beschreibung im Text).

dem Musterbild erstellt. Dies ist schon in Abschnitt 5.1.9 beschrieben worden.

Danach wird eine Gruppe Ω von neuen Schichten hinzugefügt. Jede neue Schicht l dieser Gruppe ist genau einem Konzept $o(l)$ der später zu erkennenden Konzepte O zugeordnet.

Eine Schichtengruppe wird neu erstellt, trainiert und assoziiert. Das geschieht iterativ solange, bis der Erkennungsfehler ε^L der Trainingsmuster unter einem Schwellwert ϑ^L liegt. Ist diese Bedingung erfüllt, wird überprüft, wie gut die Erkennung auf der Menge Q^T von Testmustern funktioniert. Liegt deren Erkennungsfehler ε^T über einem Schwellwert ϑ^T, werden wieder neue Schichtengruppen eingeführt. Ist die letzte Bedingung erfüllt, terminiert das Verfahren. Für den Einsatz in realen Anwendungen kann das Provadero-Verfahren mit verschiedenen Analysestufen ausgestattet werden, die im Abschnitt 5.3 beschrieben werden. Das Konzept $o^{det}(\Omega, \psi)$ ist das Ergebnis eines exklusiven Klassifikators, wie er in Abschnitt 5.3.1 beschrieben ist.

Die Fehlerraten werden wie folgt berechnet:

$$\varepsilon^L(\Omega, \psi) = 1 - \frac{\sum_{q^L} \mathcal{I}\big(o(q^L), o^{det}(\Omega, \psi)\big)}{|Q^L|}, \tag{5.81}$$

$$\varepsilon^T(\Omega, \psi) = 1 - \frac{\sum_{q^T} \mathcal{I}\big(o(q^T), o^{det}(\Omega, \psi)\big)}{|Q^T|} \quad \text{mit} \tag{5.82}$$

$$\mathcal{I}(a, b) = \begin{cases} 1 & : \text{ für a=b} \\ 0 & : \text{ sonst} \end{cases} \tag{5.83}$$

Die Funktion $o(q)$ liefert das Konzept, das einem Muster q zugeordnet ist.

Das Lernen einer Schichtengruppe Ω ist in verschiedene Teilschritte unterteilt, die schon in Abbildung 5.12 gezeigt wurden. Nach jedem Lernschritt wird der zugehörige Assoziationsschritt durchgeführt, um Daten für den nächsten Lernschritt bereitzustellen. Die Lernschritte werden in den folgenden Abschnitten beschrieben.

Die Freigabe $\mathcal{E}$ beschreibt, welche Komponenten unterliegender Schichten besonders geeignet sind, Konzepte zu unterscheiden. Dies erfolgt anhand einer Freigabematrix E. Diese wird mit dem in Abschnitt 5.2.1 beschriebenen Verfahren so eingestellt, dass nur Komponenten verwendet werden, bei denen sich für die Bilder eines Konzeptes möglichst hohe und für unterschiedliche Konzepte möglichst niedrige Konfidenzen ergeben.

Die Rücktransformation $\mathcal{B}$ soll Varianzen ausgleichen. Die Rücktransformationsparamter $\mathbf{b}^o$ und $\mathbf{b}^s$ sollen mit dem in Abschnitt 5.2.2 beschriebenen Verfahren so eingestellt werden, dass sich für unterschiedliche Abbildungen eines Konzeptes möglichst identische Zustandsvektoren ergeben.

Als Vorverarbeitungsschritt für die Projektion, werden mit der Skalierung $\mathcal{S}$ die verwendeten Komponenten mittelwertfrei gemacht und dann auf eine einheitliche Varianz skaliert. Die dazu notwendigen Momentvektoren $\mathbf{m}^1$ und $\mathbf{m}^2$ werden mit dem im Abschnitt 5.2.3 beschriebenen Verfahren bestimmt.

Eine Schichtengruppe enthält für jedes zu lernende Konzept $o \in O$ genau eine Schicht. Diese soll spezifisch auf ein Teilkonzept sein, das einen wesentlichen Baustein zur Repräsentation dieses Konzeptes darstellt. Sie soll auch möglichst invariant auf Varianzen dieses Teilkonzeptes reagieren. Das wird mit dem in Abschnitt 5.2.4 beschriebenen Verfahren zum Lernen der Spezifitäten $\mathbf{s}$ gewährleistet.

5.2.1 Lernen der Freigabe

Das Lernen der Freigabe wird für jede Schicht l durchgeführt und ist dann für alle Muster gleichermaßen gültig. Die zu bestimmende Freigabematrix E soll bei der Assoziation nur solche Komponenten x_m^Ω der Zustandsvektoren zum Auswerten weitergeben, die ein gutes Unterscheiden der Konzepte ermöglichen. Für jede Komponente wird ein Wert η_m eingeführt, der beschreibt, wie gut die Komponente diesem Anspruch genügt. Er wird berechnet aus dem Verhältnis der mittleren Aktivität dieser Komponente bei Mustern des zu detektierenden Konzeptes gegenüber der mittleren Aktivität der Komponente bei anderen Mustern.

$$\eta_m(l) = \frac{c_m^o(l)}{c_m^r(l)} \tag{5.84}$$

Der Wert $\mathbf{c}^o(l)$ beschreibt die gemitteltete Aktivierung von Mustern die das gleiche Konzept haben, wie das der Schicht zugeordnete. Der Wert $\mathbf{c}^\rho(l)$ beschreibt die gemitteltete Aktivierung von Mustern, die ein anderes Konzept haben, wie das der Schicht zugeordnete. Berechnet werden sie wie folgt:

$$c_m^o(l) = \frac{1}{w^o} \sum_{q^L \in Q^L | o(q^L) = o(l)} \sum_p c_m^p(q^L, l, p), \tag{5.85}$$

$$c_m^\rho(l) = \frac{1}{w^\rho} \sum_{q^L \in Q^L | o(q^L) \neq o(l)} \sum_p c_m^p(q^L, l, p). \tag{5.86}$$

Wobei die w-Faktoren eine Normierung über die aufsummierten Rasterpositionen vornehmen:

$$w^o = \sum_{q^L \in Q^L | o(q^L)=o(l)} \sum_p 1, \tag{5.87}$$

$$w^\rho = \sum_{q^L \in Q^L | o(q^L)\neq o(l)} \sum_p 1. \tag{5.88}$$

Die errechneten η_m können der Größe nach sortiert werden. Die Komponenten, die das gewünschte Konzept am besten von anderen unterscheiden, haben dann hohe Werte. Zur Auswahl der Komponenten kann man nun noch einen Schwellwert vorgeben oder wie in der hier gewählten Realisierung eine maximale Gesamtanzahl N von Komponenten, die für jede Schicht mit den am besten unterscheidenden Komponenten aufgefüllt wird. Bei der hier gewählten Lösung wird für vektorkodierende Komponenten zusätzlich sichergestellt, dass sobald eine Komponente Verwendung findet, die andere auch freigegeben wird. Die Bestimmung von E ist einfach. Es wird zunächst eine Identitätsmatrix mit $M(\Omega) \times M(\Omega) \mid l \in \Omega$ erzeugt. Aus dieser werden dann die Zeilen der korrespondierenden nicht verwendeten Komponenten gelöscht. Es ergibt sich eine $N(\Omega) \times M(\Omega) \mid l \in \Omega$ Matrix.

5.2.2 Lernen der Rücktransformationsparameter

Mit diesem Verfahren sollen die Rücktransformationsparameter $\mathbf{b}^o$ und $\mathbf{b}^s$ eingestellt werden. Dafür wurde noch keine konkrete Realisierung entwickelt. Es wird nur das zugrunde liegende Prinzip beschrieben: Unterschiedliche Muster eines Konzeptes sollen zu ähnlichen Assoziationen führen. Dies soll so realisiert werden, dass sich für gleiche v-Werte von Mustern eines Konzeptes auch gleiche Ausprägungen v für die Folgeschicht ergeben.

$$\forall o \in O,\, q_1^L,\, q_2^L \in Q^L \wedge o = o(q_1^L) = o(q_2^L) \wedge q_1^L \neq q_2^L : \\ \sum_p x_n^b(l, p, q_1^L) \stackrel{!}{=} \sum_p x_n^b(l, p, q_2^L). \tag{5.89}$$

5.2.3 Einstellen der Skalierung

Die Momente $\mathbf{m}^1$ und $\mathbf{m}^2$ sollen die Zustandsvektoren einer Schicht l mittelwertfrei machen und skalieren. Gleichung (5.67) beschreibt die zugehörige Berechnung. Es sollen jeweils alle Zustandsvektoren berücksichtigt werden, die in Mustern q^L enthalten sind, deren Konzept $o(q^L)$ dem der Schicht

$o(l)$ entspricht. Zur Berechnung von $\mathbf{m}^1$ werden die Zustandsvektoren mit Ihrer Konfidenz gewichtet und dann gemittelt. Für jede Vektorposition n wird dann noch mit dem Faktor w_n auf die Konfidenz normiert. Die Berechnung von $\mathbf{m}^2$ bildet zunächst die Differenz zum Mittelwert und quadriert dann das Ergebnis: Es werden alle Varianzen des zu l gehörigen Konzeptes betrachtet.

$$m_n^1(l) = \frac{1}{w_n(l)} \sum_{q^L \in Q^L | o(l)=o(q^L)} \sum_p c_n^b(q^L,l,p)\, x_n^b(q^L,l,p), \quad (5.90)$$

$$m_n^2(l) = \frac{1}{w_n(l)^2} \sum_{q^L \in Q^L | o(l)=o(q^L)} \sum_p c_n^b(q^L,l,p)\Big(x_n^b(q^L,l,p) - m_n^1(l)\Big)^2 \quad \text{mit} \quad (5.91)$$

$$w_n(l) = \sum_{q^L \in Q^L | o(l)=o(q^L)} \sum_p c_n^b(q^L,l,p). \quad (5.92)$$

5.2.4 Lernen der Spezifität

Für eine neue Schichtengruppe wird für jede Schicht die Spezifität $\mathbf{s}(l)$ berechnet. Sie bestimmt hauptsächlich, wie sich die Schicht bei einer Assoziation verhält. Die Berechnung soll so erfolgen, dass $\mathbf{s}(l)$ in Richtung der größten Varianz der Zustandsvektoren des l zugeordneten Konzeptes $o(l)$ zeigen soll. Dazu soll eine Hauptkomponentenanalyse vorgenommen werden. Zur Berechnung wird das „Hebbian-Based Maximum Eigenfilter"-Verfahren von Oja (1982) angewendet, welches die erste Hauptkomponente der Zustandvektoren-Verteilung berechnet. Das Verfahren arbeitet iterativ und addiert die einzelnen Zustandsvektoren in die Richtung an die Spezifität, die durch die Projektion mit der bisherigen Spezifität gegeben ist. So bewegt sich die Spezifität in die Richtung mit der größten Varianz. Die Hilfsvariable τ zählt über alle Rasterpunkte p aller Trainingsmuster q^L des zugehörigen Konzeptes:

`wiederhole für alle` $p,\ q^L \in Q^L$: (5.93)

$$\mathbf{s}^*(l,\tau+1) = \begin{cases} \mathbf{s}^*(l,\tau) + \bar{c}^s(q^L,l,p)\,\mathbf{x}^s(q^L,l,p) \ : \ \mathbf{s}^*(l,\tau)^T\,\mathbf{x}^s(q^L,l,p) \geq 0 \\ \mathbf{s}^*(l,\tau) - \bar{c}^s(q^L,l,p)\,\mathbf{x}^s(q^L,l,p) \ : \ \mathbf{s}^*(l,\tau)^T\,\mathbf{x}^s(q^L,l,p) < 0 \end{cases}$$

`inkrementiere` τ `mit 1`

mit $\bar{c}^s(q^L,l,p) = \frac{1}{N(l)} \sum_n^{N(l)} c_n^s(q^L,l,p)$, $\mathbf{s}^*(l,0) = \mathbf{0}$, $o(q^L) = o(l)$. (5.94)

Mit $\bar{c}^s$ wird die Konfidenz des Zustandsvektors gemittelt. Die noch nicht normierte Spezifität $\mathbf{s}^*$ wird abschließend zur Spezifität $\mathbf{s}$ als Einheitsvektor normiert:

$$\mathbf{s}(l) = \frac{\mathbf{s}^*(l, \tau_{max})}{||\mathbf{s}^*(l, \tau_{max})||}. \tag{5.95}$$

5.3 Analyse

Mit dem Provadero-Verfahren lassen sich unterschiedliche Schnittstellen zu Anwendungen realisieren. Im einfachsten Fall kann man einen Exklusivklassifikator realisieren, der angibt, welches der gelernten Konzepte am wahrscheinlichsten in einem Eingabebild ψ vorhanden ist. Die Klassifikation wählt also stets eins der in O gegebenen Konzepte aus. Ein allgemeiner Klassifikator kann für jedes Eingabebild alle Konzepte aus O zuordnen. Es ist auch ein Konzeptlokalisierer möglich, der angibt, an welcher Position im Eingabbild ein Konzept gegeben ist. Die zugehörige Region kann segmentiert werden.

5.3.1 Exklusiver Klassifikator

Dieser Klassifikator liefert das Konzept, was am wahrscheinlichsten im Muster q abgebildet ist: $o^{det}(\psi)$. Dazu werden die Aktivierungen der Schichten in der höchsten Schichtengruppe Ω^{top} ausgewertet. Das Resultat ist das Konzept, welches der am stärksten aktivierten Schicht zugeordnet ist:

$$o^{det}(q) = o\big(\arg\max_{l \in \Omega^{top}} \sum_p c^v(q, l, p)\big). \tag{5.96}$$

5.3.2 Allgemeiner Klassifikator

Für jede Schicht der obersten Schichtengruppe Ω^{top} wird überprüft, ob sie stärker aktiviert ist als der Schwellwert ϑ^K. In diesem Fall ist das zugehörige Konzept detektiert. Die detektierten Konzepte bilden dann die Detektionsmenge O^{det}:

$$o^{det}(q, o) = \begin{cases} o & : \text{ für } 1/|p| \sum_p c^v(q, l(\Omega^{top}, o), p) > \vartheta^K \\ \emptyset & : \text{ sonst} \end{cases} \tag{5.97}$$

$$O^{det}(q, O) = \left\{ o_1^{det}(q, o_1 \in O), o_2^{det}(q, o_2 \in O), ... \right\} \tag{5.98}$$

Die Funktion $l(\Omega, o)$ liefert die Schicht in der Schichtengruppe Ω, die dem Konzept o zugeordnet ist.

5.3.3 Konzeptlokalisierer

Für diesen Detektortyp muss eine Funktion eingeführt werden, die eine Konzeptregion definiert. Dazu muss eine zweidimensionale Struktur beschrieben werden, die eine bestimmte Aktivierung enthalten muss. Mit dieser Funktion kann man dann Bereiche eines Konzepts und eine zugehörige Konfidenz detektieren. Eine zusätzlich zu definierende Funktion kann aus dem Bereich eine diskrete Position errechnen; beispielsweise kann geometrisch gemittelt werden.

Es ist auch möglich, für ein derartig detektiertes Konzept, eine Segmentierung im Eingabebild vorzunehmen. Dazu kann untersucht werden, welche Regionen der unterliegenden Schicht einen Beitrag zur Assoziation des Konzeptes geleistet haben. Dies kann rekursiv für alle unterliegenden Schichten erfolgen. Ein ähnliches Verfahren wurde schon erfolgreich zum Segmentieren in einer hierarchischen Schichtenarchitektur angewendet (Teichert and Malaka, 2003). Mit den Provadero-Komponenten und den Lernverfahren werden sich subsymbolisch repräsentierende Teilkonzepte ergeben. Diese werden so typischerweise nicht von Menschen verwendet und können daher auch nicht semantisch zugeordnet werden.

5.4 Zusammenfassung und Bewertung

Es sind Realisierungen für alle Module des Provadero-Verfahrens auf Kapitel 4 vorgestellt worden. Die Module konnten so realisiert werden, dass nur relativ wenig Konfigurationsparameter notwendig sind.

Die Laufzeit- und Speichergrößenordnungen sind dabei linear in der Anzahl der Muster, linear in der Anzahl der Pixel, linear in der Anzahl der benötigten Schichtengruppe und linear in der Anzahl der zu lernenden Konzepte. Doch die Anzahlen multiplizieren sich zum Gesamtaufwand. So können relativ schnell unhandliche Größen erreicht werden, wenn viele Muster in hohen Auflösungen mit vielen Konzepten gelernt werden sollen.

Insbesondere die lineare Abhängigkeit des Speicheraufwandes in der Anzahl der Muster sorgt dafür, dass schon mit kleinen Mustersätzen Grenzen erreicht werden. Andere Verfahren verwenden meist sequenzielle Lernverfahren und haben dadurch keinen großen Speicheraufwand. Auch für das Provadero-Verfahren sind solche Techniken denkbar. Sobald auf einer Schicht l die Spezifität $\mathbf{s}(l)$ gelernt wurde, könnte man viele Zwischenergebnisse verwerfen. Es müssten lediglich die Zustandsvektoren x^l für jede Schicht gespeichert werden, damit höhere Schichten noch in Abhängigkeit

ihrer Freigabefunktion darauf zugreifen können. Die Ordnung des Speicheraufwands verringert sich dadurch aber noch nicht. Dies wäre erst möglich, wenn man das Lernen selbst sequenzieren könnte. Schwierig wird das für die Freigabefunktion und die Skalierung, die dann nicht mehr über alle Muster ausgewertet werden könnten. Spätere Erweiterungen realisieren möglicherweise eine gemischte vorwärts- und rückwärtsgerichtete Verarbeitung. Dies würde die Sequenzierung des Lernverfahrens zusätzlich erschweren.

Ist das Lernen erfolgt, müssen nur noch die Parametrisierungen der einzelnen Module gespeichert werden. Das Verfahren ist dann nur noch linear abhängig von der Anzahl der Pixel, Schichtengruppen und Konzepte.

Während das Sequenzieren des Provadero-Lernverfahrens grundlegende Änderungen erfordert, läßt sich das Verfahren sehr gut parallelisieren, weil der Großteil des Datenaustausches lokal und regional stattfindet.

Kapitel 6

Evaluierung

In diesem Kapitel werden Simulationen des Diffusionsverfahrens für Strukturwerte und der Provadero-Realisierung dargestellt, diskutiert und abschließend bewertet. Zunächst wird das Diffusionsverfahren für Strukturwerte untersucht und gezeigt, dass es invariant gegenüber Positions-, Skalierungs- und Rotationstransformationen sind.

Für das Provadero-Gesamtsystem wird anschließend gezeigt, dass es geeignet ist, das Erkennen von Objekten unter klassischen Varianzen, insbesondere aber auch das Erkennen unter abstrakten Varianzen zu ermöglichen.

6.1 Diffusionsverfahren

6.1.1 Evaluationsstrategie

Da das Diffusionsverfahren zur Interpolation in der zur Evaluation verwendeten Zusammenstellung von Provadero-Komponenten nicht benötigt wird, werden an dieser Stelle auch keine Simulationen dazu beschrieben. Bisher durchgeführte Simulationen zeigen aber, dass das Verfahren erwartungsgemäß funktioniert und die in Abschnitt 5.1.2 beschriebenen Anforderungen erfüllt.

Das Diffusionsverfahren für Strukturwerte besitzt als Parameter lediglich die Anzahl der Iterationen, die für die Diffusion durchgeführt werden. Die Evaluation des Verfahrens ist daher einfach zu realisieren. Als Eingabebilder werden zunächst einfache Helligkeitsübergänge getestet. Danach werden Untersuchungen mit transformierten Bildinhalten angestellt und schließlich werden typische Realweltbilder der Forschungsgemeinschaft zum Bildverarbeiten und -verstehen untersucht. Abschließend werden die Ergebnisse diskutiert.

6.1.2 Einstufiger gerader Übergang

Abbildung 6.1 zeigt Simulationen auf einem einfachen Hell-Dunkel-Übergang in zwei Dimensionen. In der ersten Spalte sind die Konfidenzwerte zu sehen. Eine höhere Konfidenz ergibt ein helleren Bildwert. In der zweiten Spalte sind die Vektorinformationen von Gradient und Translation dargestellt. Die Helligkeit kodiert die Vektorlänge und die Farbe kodiert die Vektorrichtung. Die Farbe Cyan beschreibt einen nach rechts zeigenden und die Farbe Grün einen nach oben zeigenden Vektor, wie man in Abbildung 6.1 b) einfach nachvollziehen kann. Andere Farbwerte ergeben sich entsprechend eines umgebenden Farbkreises eines HSV-Farbmodells. Die oberste Zeile (a) von Abbildung 6.1 zeigt das Eingabebild (mitte) und dessen Konfidenz (links), die volle Aktivität über den gesamten Eingabebildbereich hat. Der gesamte Eingabebildbereich ist also „sicher“ gegeben. Das kombinierte Bild (rechts) zeigt das Eingabebild multipliziert mit der Konfidenz, was in diesem Fall wieder das Eingabebild ergibt. Die nächste Zeile (b) zeigt den Gradient $\hat{\mathbf{g}}$ und die zugehörige Konfidenz $\hat{c}$ auf den Zellgrenzen. Die Aktivierungen haben in diesem Fall eine Breite von zwei Pixeln, was bei einem einfachen Helligkeitsübergang und dessen Filterung mit Gleichung (5.13) plausibel ist. Die Bilder von den Grenzaktivierungen werden nur einmal dargestellt, weil sie über die Iterationen konstant bleiben. Dies ist plausibel, weil sich durch das Diffusionsverfahren in der Umgebung nie größere Werte c^g, als die Quellwerte c_b^h ergeben und diese nach Gleichung (5.20) und (5.24) stets dominant bleiben.

Für die Bilder werden Zwischenergebnisse bei 30 und 300 Iterationen dargestellt. So kann man erkennen, wie sich die Werte darstellen, wenn sie sich durch die Diffusionsiterationen noch nicht über das gesamte Bild ausbreiten konnten und im zweiten Fall kann man das Ergebnis der Ausbreitung über das ganze Bild sehen. Zusätzlich sieht man, ob eine hohe Iterationsanzahl zu Sättigungserscheinungen oder anderen Artefakten führt.

Die Zeilen (c) und (d) zeigen die Gradienteninformation und deren Konfidenz. Die Kombination wurde wieder anhand einer Multiplikation errechnet. Man sieht, wie sich in den Regionen der Übergänge homogene Beschreibungen deren Richtungen ergeben. Wo sie aufeinander stoßen, findet keine Vermischung statt, weil in diesem Beispiel durch die Symmetrie des Eingabebildes gleiche Konfidenzen erzeugt werden und daher kein Fluss mit Gradienteninformation aus der Umgebung stattfindet. In der Analogie der Wasserbehälter würden zwei Behälter mit unterschiedlichen Farben nebeneinander stehen. Es würde aber kein Austausch stattfinden, weil die Behälter die gleiche Höhe haben. Die Außenecke der Helligkeitsspünge erzeugt einen Bereich, der sich wie um eine Rundung ergibt (siehe nächsten

Abschnitt). Die Richtung des Gradienten zeigt stets auf die Außenecke.

Die Zeilen (e) und (f) zeigen Translationsinformation und deren Konfidenz. Man sieht gut, wie sich die Länge der Vektorinformation zu den Außenrändern hin aufbaut und in die Richtung des Quellgradienten zeigt. Die Randwerte des Bildes nach 30 Iterationen sind genauso groß, wie die im Bild nach 300 Iterationen an den gleichen Positionen. Die Bilder sind lediglich zur besseren Sichtbarkeit unterschiedlich skaliert. Die Konfidenz ist nach erfolgter Diffusion recht homogen verteilt. Lediglich direkt neben den Gradientenquellen sind sie etwas erhöht und in den Bereichen, wo unterschiedliche Gradienten aufeinander stoßen, kommt es zur Auslöschung der Konfidenz. Dieser Effekt ist darin begründet, das die Gradientenvektoren im Skalarprodukt von Gleichung (5.29) an diesen Positionen unterschiedliche Richtungen haben.

6.1.3 Einstufiger runder Übergang

Ergebnisse von Diffusionssimulationen auf einem runden Übergang sind in Abbildung 6.2 dargestellt. Es gelten dieselben Erörterungen, wie im vorigen Abschnitt. Zusätzlich sieht man an den konzentrisch ausgerichteten Vektorstrukturen, dass das Verfahren nahezu rotationssymmetrisch funktioniert und auch kaum Artefakte durch relativ grob verpixelte Rundungen entstehen, die in dem Eingabebild ohne Grauwertinterpolation gegeben sind.

6.1.4 Mehrstufiger Übergang

Abbildung 6.3 zeigt Diffusionssimulationen auf einem Mehrfachübergang. Die Gradienteninformation bildet sich wie bei dem einfachen Übergang homogen um die Quellregion. Treffen Gradienten mit unterschiedlicher Richtungsinformation aufeinander, stoppt die Ausbreitung, wenn gleiche Konfidenzniveaus erreicht sind. Die Translationsinformation berechnet sich korrekt für die einzelnen Teilbereiche. Die Simulationen zeigen auch für dieses Experiment stabile Wertausprägungen über Diffusionsiterationen.

6.1.5 Positions-, Skalierungs- und Rotationsinvarianz

Ein Verfahren zur Analyse von zweidimensionaler Kontextinformation in Bildern sollte invariant sein gegenüber Transformationen des Eingabebildes bezüglich Position, Skalierung und Rotation. Das soll dieses Experiment zeigen, indem entsprechend transformierte Teilbildbereiche untersucht werden, die den gleichen Inhalt zeigen. Abbildung 6.4 zeigt, das diese Invarianzen

gegeben sind. Die regionalen Strukturen von Gradient und Translation sind innerhalb der Objekte stets identisch.

6.1.6 Realweltbilder

Abschließend sollen noch Diffusionssimulationen auf zwei Realweltbildern durchgeführt werden, die sich als Standard zur Beurteilung von Bildverarbeitungsverfahren etabliert haben (Abbildung 6.5 und 6.6). Die Bilder zeigen interessante Bildstrukturen im Vorder- und Hintergrund sowie verschiedene Übergänge und Störungen. Der Gradient auf den Zellgrenzen entspricht dem Ergebnis, das man mit einem klassischen Sobelfilter erhalten würde. Daher können die Bilder in den Zeilen (b) zum Vergleich verwendet werden. Man sieht, dass der Sobelfilter lokal wirkt, während das Diffusionverfahren stark ausgeprägten Gradienten größere und zusammenhängendere Regionen zuordnet. Wie man gut in Abbildung 6.6 in den Zeilen (c) und (d) sehen kann, wird auch das Bildrauschen im Bereich des Himmels stark vermindert.

6.1.7 Bewertung

Das Diffusionsverfahren für Strukturwerte funktioniert nach den Vorgaben aus Abschnitt 5.1.3. Es sind keine Artefakte aus Fehlfunktionen erkennbar. Das Verfahren eignet sich, um Kontextinformation aus zweidimensionalen Signalverteilungen zu extrahieren. Das Verfahren ist auch isoliert von den restlichen Provadero-Modulen einsetzbar und kann als Strukturextraktionsverfahren genutzt werden.

Gegenüber klassischen Filtern steht beim Provadero-Diffusionsverfahren Kontextinformation nicht nur innerhalb der meist kleinen Filtermasken bereit, sondern für den gesamten Bildbereich Die Kontextinformation des Diffusionsverfahren beschreibt den dominierenden Gradienten in der Umgebung - auch wenn dieser weit entfernt ist. Das ist insbesondere für Assoziationssysteme von Vorteil, die räumlich verteilte Information auswerten müssen. Da die Repräsentationen als Vektorinformation gegeben sind, können sie sehr leicht skaliert oder rotiert werden. Das ist vorteilhaft, wenn man in den Repräsentationen regionale Verzerrungen ausgleichen will, die sich z.B. durch perspektivische Projektion bei der Bildaufnahme ergeben haben.

Man kann einwenden, dass das Verfahren bei sehr kleinen lokalen Bildstrukturen/Texturen keine regionale Information bereitstellt. Das hat man mit klassischen Filtern jedoch auch nicht. Diese Bereiche zusammenzufas-

sen, muss die Aufgabe eines Assoziationssystems sein, welches von diesen lokalen Inhalten auf regionale abstrahieren kann.

Für das Verfahren muss nur ein Parameter eingestellt werden: die Anzahl der Iterationen. Dabei ist selbst das unkritisch, denn wenn das Verfahren erstmal die Strukturinformation über das Bild verteilt hat, bleiben die Strukturen stabil. Man kann also das Einstellen dieser Größe leicht automatisieren, in dem man die Anzahl in Abhängigkeit der größten Bilddimension einstellt.

Es lassen sich leicht zusätzliche Parameter einführen. Z.B. könnte man die Terme in den Gleichungen (5.12) unterschiedlich gewichten. Das verändert jedoch hauptsächlich nur die Geschwindigkeit, mit der sich die Diffusion über die Iterationen ausbreitet. Teichert and Malaka (2006) haben für ein ähnliches Verfahren festgestellt, dass eine solche Gewichtung weite Bereiche der Einstellmöglichkeiten stabile Ergebnisse liefert.

Der Rechenaufwand ist jeweils linear in der Anzahl der Bildpixel und der Anzahl der Iterationen. Der Aufwand zum Berechnen von Sätzen mit klassischen Maskenfunktionen ist nicht direkt vergleichbar. Allerdings werden diese schnell sehr aufwändig, wenn Sätze von Masken in unterschiedlichen Richtungen und Skalierungen gerechnet werden sollen.

Das Diffusionsverfahren kann auch verwendet werden, wenn sich die Inhalte des zu analysierenden zweidimensionalen Signals ändern (Teichert and Malaka, 2006). Das ist insbesondere dann interessant, wenn Repräsentationen durch eine Bottom-Up als auch durch eine Top-Down-Verarbeitung verändert werden. Das kann zum Beispiel durch eine Top-Down-Hemmung von Repräsentationen erfolgen, die Bottom-Up erstellt wurden und noch Mehrdeutigkeiten in sich tragen. Die Strukturinformationen müssen nicht für eine geänderte Repäsentation vollständig iterieren. Das Ändern von Repräsentationen und das Diffusionsiterieren kann auch „Zug um Zug“ durchgeführt werden.

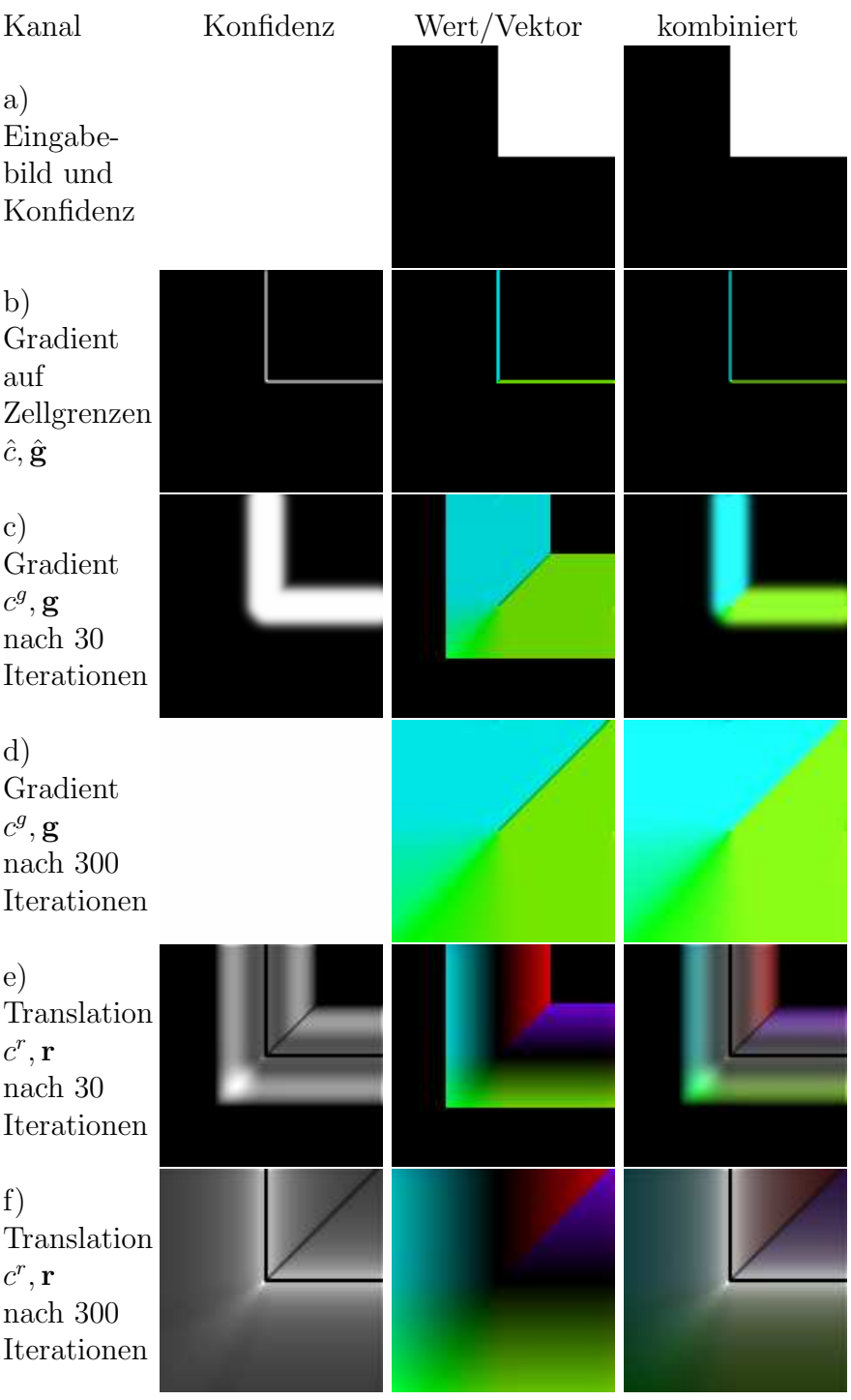


Abbildung 6.1: Diffusionsexperiment auf einem geraden Hell-Dunkel-Übergang in zwei Dimensionen mit 128x128 Pixeln (siehe Text).

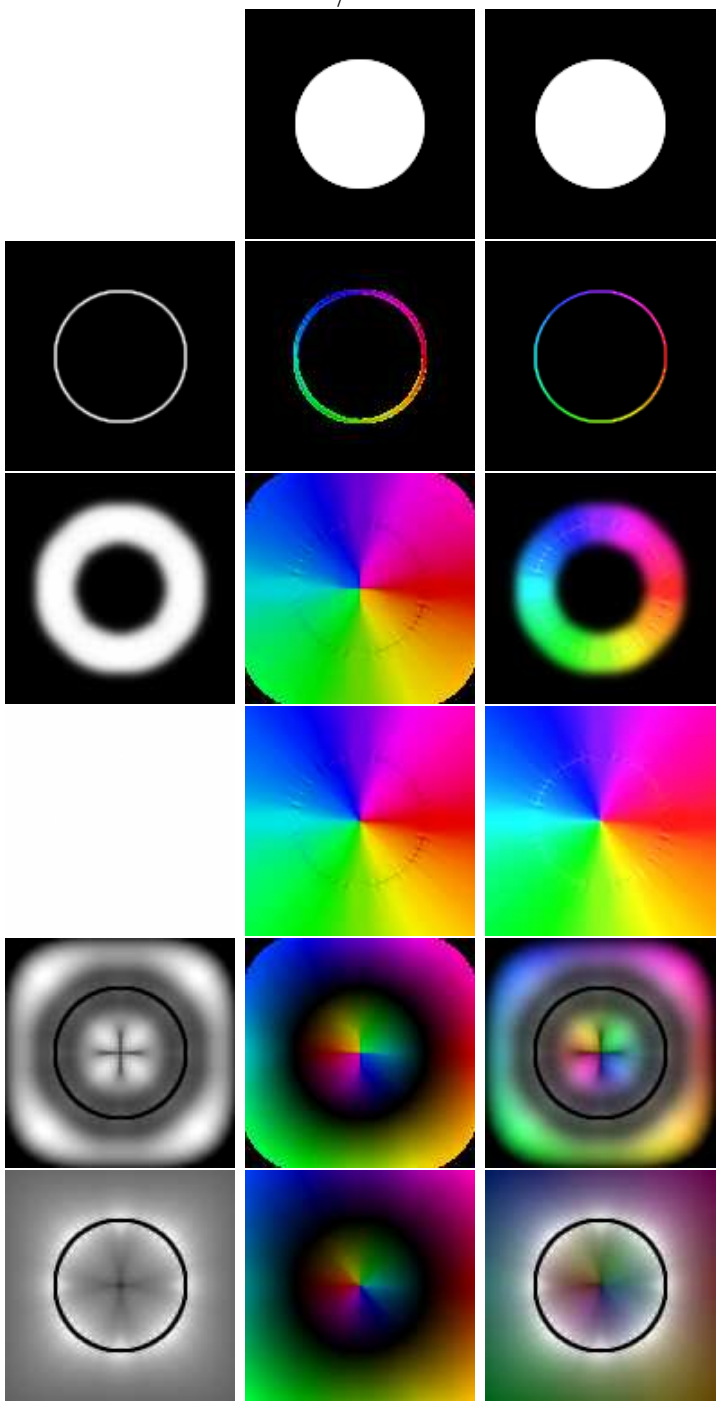


Abbildung 6.2: Diffusionsexperiment auf einem runden Hell-Dunkel-Übergang in zwei Dimensionen mit 128x128 Pixeln (siehe Text).

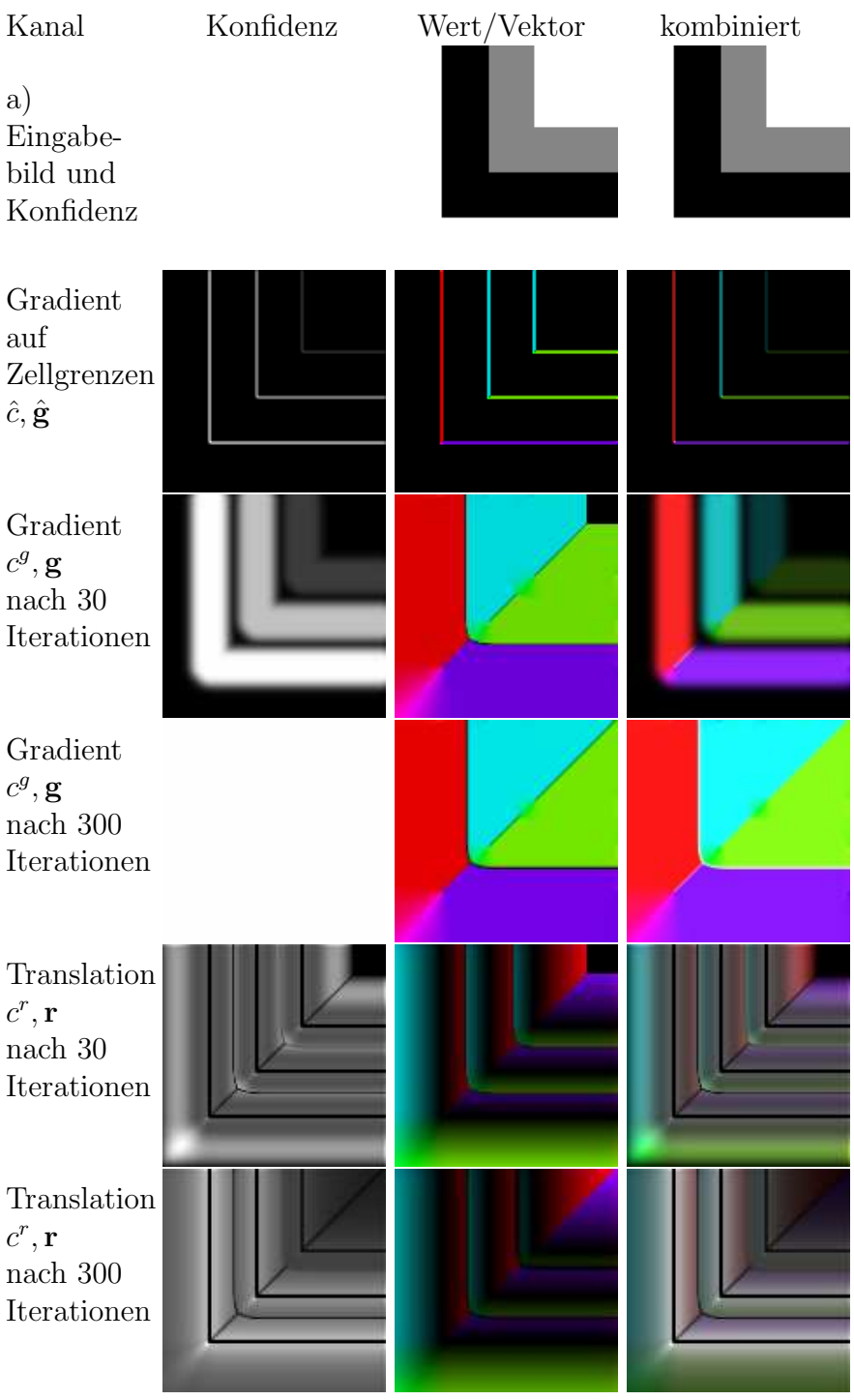


Abbildung 6.3: Diffusionsexperiment auf einem abgestuften Hell-Dunkel-Übergang in zwei Dimensionen mit 128x128 Pixeln (siehe Text).

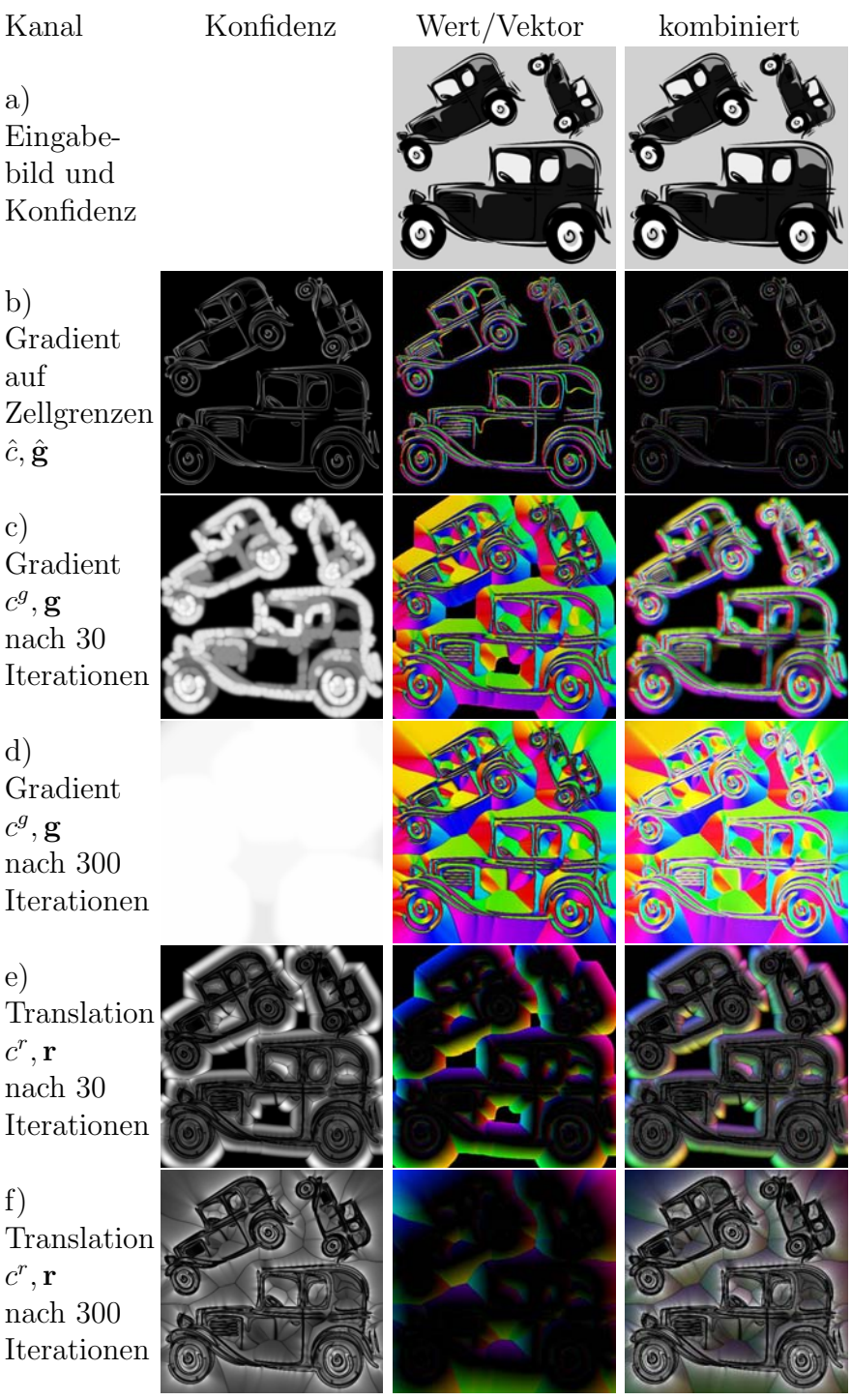


Abbildung 6.4: Diffusionsexperiment zur Positions-, Skalierungs-, und Rotationsinvarianz 512x512 Pixeln (siehe Text).

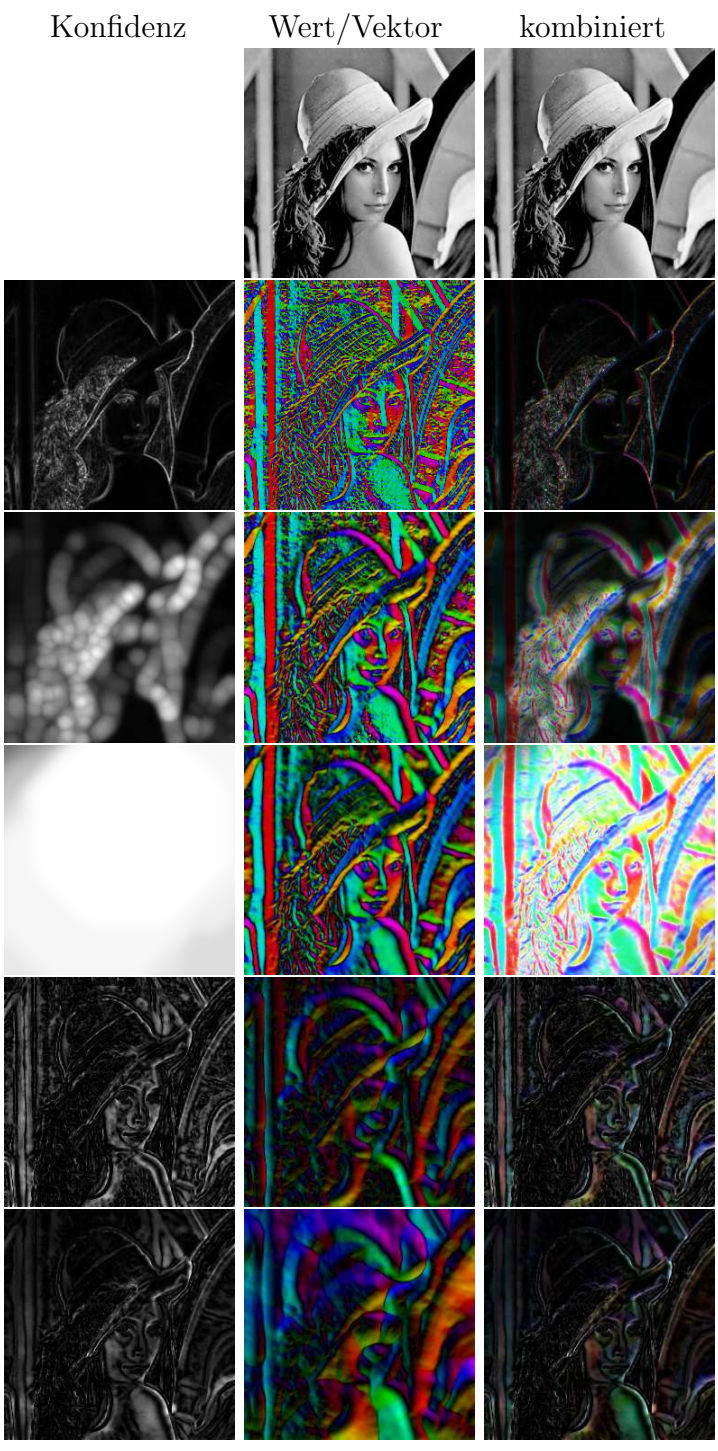


Abbildung 6.5: Diffusionsexperiment auf Realweltbild 1 mit 256x256 Pixeln (siehe Text).

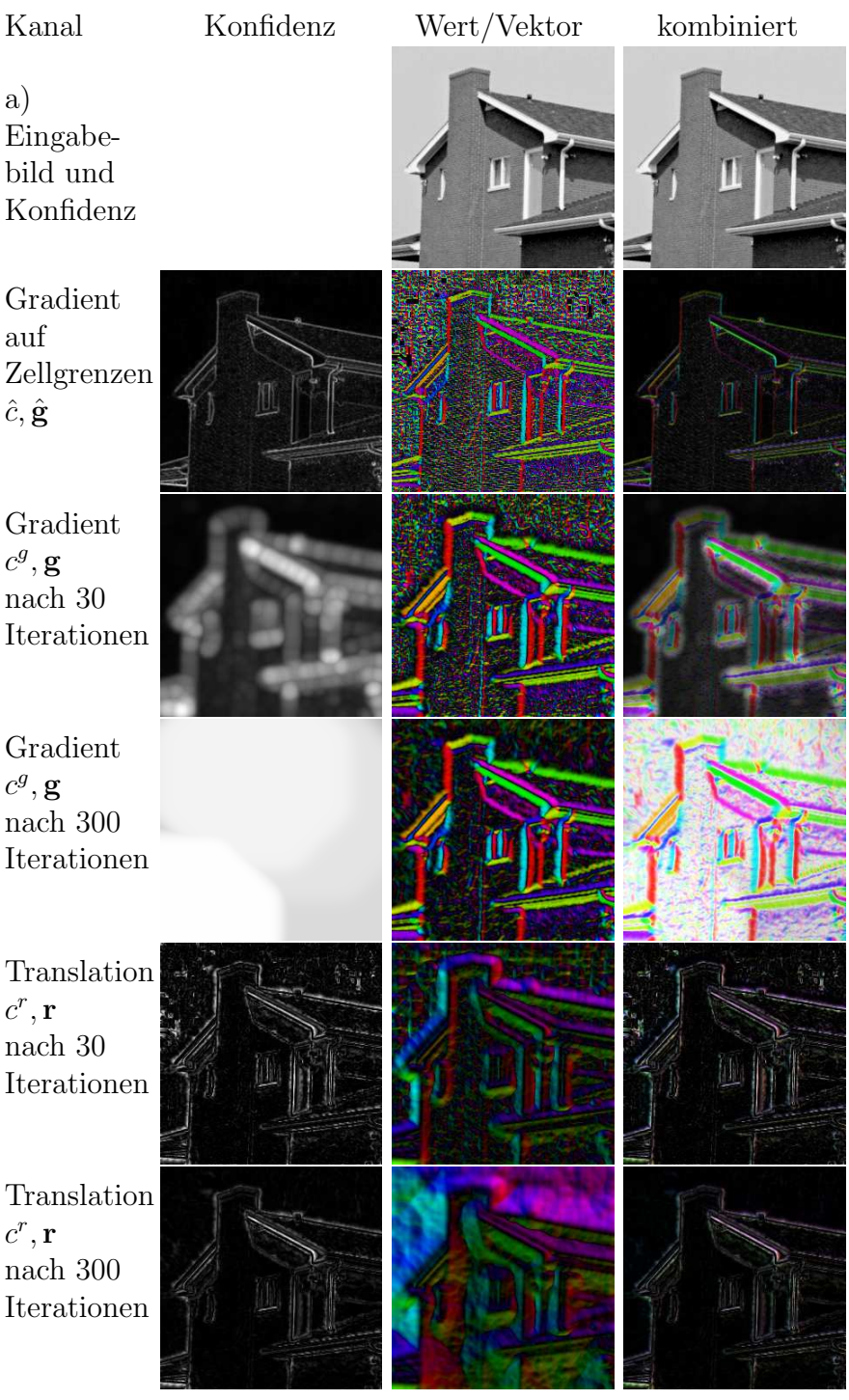


Abbildung 6.6: Diffusionsexperiment auf Realweltbild 2 mit 256x256 Pixeln (siehe Text).

6.2 Objekterkennung

6.2.1 Evaluationsstrategie

In diesem Abschnitt werden die in Kapitel 5 beschriebenen Provadero-Komponenten im Verbund evaluiert. Es soll zunächst geprüft werden, ob grundlegende Eigenschaften gegeben sind, die zur Objekterkennung notwendig sind. Später wird dann untersucht, wie das Verfahren mit abstrakten Varianzen funktioniert. Schließlich werden Simulationen auf Nachrichtenbildern vorgenommen, die Personen in unterschiedlichen Posen zeigen.

Die grundlegenden Eigenschaften werden anhand von Binärwertbildern untersucht, die geeignet sind, zu zeigen, ob das Verfahren für identische Trainings- und Testmuster auch identische Ergebnisse erzeugt. Ebenso wird untersucht, ob die Ergebnisse gleich sind, wenn identische Muster an unterschiedlichen Positionen, in unterschiedlichen Skalierungen oder unter unterschiedlichen Rotationen gegeben sind.

Mustersätze

Zum Evaluieren von Bildverarbeitungsverfahren gibt es für verschiedene Themen etablierte Bildersätze, die zum Vergleich mit anderen Verfahren verwendet werden können. Für Spezialthemen werden meist neue Bildersätze erstellt, die es ermöglichen, die besonderen Eigenheiten der Verfahren herauszustellen. Aus diesem Grund sind sie auch nicht unbedingt zum Vergleich mit anderen Verfahren geeignet. Für das Erkennen von Bildkonzepten in unterschiedlichen Ausprägungen gibt es schon einige Bildersätze (z.B Everingham et al., 2006, 2010). Für das Provadero-Verfahren werden aber dennoch neue Bildersätze erstellt. Das liegt daran, dass für die von Grund auf neu entwickelten Provadero-Komponenten zunächst grundlegende Eigenschaften überprüft werden müssen. Dabei muss darauf geachtet werden, dass nicht viele Bilder parallel verarbeitet werden können und daher ein guter Kompromiss zwischen Auflösung und der zur Verfügung stehenden Anzahl von Trainings- und Testbildern realisiert werden muss. Das gilt auch für die Bildersätze, die innerhalb des BMBF-Projektes „Bild-Film-Diskurs“ (BFD-Projekt[1]) gegeben sind und die im Rahmen dieses Beitrags verwendet werden sollen.

Der BFD-Projekt-Bildersatz enthält Nachrichtenbilder, die Personen in Posen wie Handschlag oder Eid zeigen. Es sind auch andere Posen aus Nachrichtenbildern enthalten, wie z.B. Pieta: eine Person trägt eine andere

[1] `http://dm.tzi.de/research/visual-computing/bild-film-diskurs`

in den Armen. Andere Bilder zeigen Amokläufer in sog. Selbstprotraits. Sie zielen mit ihrer Waffe in Richtung der Kamera.

Die BFD-Bilder zeigen, wie die meisten Realweltbilder eine sehr große Ausprägungsvielfalt auf hohen Abstraktionsebenen. Es gibt aber auch Gemeinsamkeiten auf weniger hohen Abstraktionsebenen. So haben z.B. die Handschlagbilder relativ ähnliche visuelle Strukturen im Bereich des Handschlags. Das gilt analog für den Eid. Die Amokläufer halten meist eine Waffe ins Bildzentrum. Bei den Pietabildern liegt die Abstraktionsebene der Gemeinsamkeiten schon wieder deutlich höher. Der BFD-Testsatz stellt allein deshalb eine Herausforderung dar, weil klassische Verfahren kaum anwendbar sind (Schäfer, 2009). Wenn Provadero also besser funktioniert, wäre das schon ein Gewinn.

Zum Testen grundlegender Mustererkennungseigenschaften wurden einige Binärbildtestsätze generiert, die später genauer beschrieben werden.

Als Auflösung wurde für alle Binärbilder eine Auflösung von 32x32 gewählt. Die Bilder des BFD-Projektes wurden mit einer Auflösung von 128 Pixeln in ihrer größten Dimension verwendet.

Zu jedem Muster ist ein Maskenbild gegeben. Das Training und die Auswertung wurde in dieser Maskenregion vorgenommen.

Konfiguration und Parametrisierung

Die in den Abschnitten 4.2.4 und 5.1.6 beschriebene Rücktransformation wird in den folgenden Simulationen nicht angewendet. Dafür gibt es zwei Gründe. Der erste Grund ist, dass die zur Verfügung stehende Hardware mit dem Speicherplatzbedarf der übrigen Provadero-Komponenten schon so ausgelastet ist, dass nur kleine Bildersätze mit geringer Auflösung gerechnet werden können und daher für ein zusätzliches Lernverfahren mit zusätzlicher Komponentenrepräsentation $\mathbf{x}^b$ kein Platz zur Verfügung steht. Der zweite Grund ist, dass die übrigen Provadero-Komponenten noch weiter optimiert werden sollten, damit die Rücktransformation auf besseren Detektionsergebnissen laufen kann. Das Unterlassen der Rücktransformation wird so realisiert, dass das Lernverfahren nicht angewendet wird und die Tranformationsparameter β_n^o und β_n^s auf 0 gesetzt werden. Praktisch kann man also $\mathbf{x}^b = \mathbf{x}^e$ und $\mathbf{c}^b = \mathbf{c}^e$ setzen.

Alle anderen Provadero-Komponenten werden wie in Abschnitt 5 beschrieben eingesetzt. Die Parametrisierung der Komponeten erfolgt mit Werten, die für Simulationen auf dem BFD-Bildersatz am besten funktionierten. Um diese zu finden, wurden die Werte nacheinander in kleinen Schritten verändert und die zugehörigen Simulationsergebnisse wurden bewertet. Danach wurden die Werte in kleinen Schritten entsprechend der Be-

wertung verändert. Diese Vorgehensweise wurde mehrfach wiederholt, bis keine Verbesserungen mehr möglich waren. Um zu vermeiden, dass man mit der Parametrisierung nur ein lokales Minimum der Ergebnisse erreicht, wurden zwischendurch auch immer wieder größere Schrittweiten ausprobiert.

Für die Anzahl N der verwendeten Vektorkomponenten der Freigabefunktion (5.47) hat sich 7 als erfolgreich herausgestellt. Es wurde noch nicht probiert, diesen Wert für jede Schichtengruppe separat zu bestimmen. Es könnte sinnvoll sein, auf höheren Schichtengruppen mehr Vektorkomponenten freizugeben, um für die Spezifitäten komplexere Abhängigkeiten realisieren zu können. Bei den Vorabsimulationen stellte sich heraus, dass der N-Parameter einigermaßen robust ist. So haben Simulationen mit 10 oder 5 Vektorkomponenten immer noch brauchbare Ergebnisse produziert. Wesentlich größer konnte der Wert nicht gewählt werden, da die Simulationen für die zur Verfügung stehende Hardware zu aufwändig geworden wären. Wird die Anzahl zu klein gewählt, verschlechtern sich die Ergebnisse. Das liegt wahrscheinlich daran, dass sich die Spezifitäten der einzelnen Schichten nicht vielfältig genug ausprägen können.

Der Projektionsparameter δ bestimmt die gaußförmige Aktivierungsfunktion (5.73). Sie bestimmt den Radius, der vorgibt, ob ein projizierter Zustandsvektor $\mathbf{x}^s$ als erkannt aktiviert wird oder nicht. Dieser Parameter ist wesentlich für das Provadero-Verfahren. Er bestimmt, wie stark die Schichten mit Aktivität gefüllt werden. Für die Ergebnisse der Vorabsimulationen hat sich ein Wert von 0,1 als erfolgreich herausgestellt. Er ergibt eine spärliche Repräsentation auf den Schichten. Dadurch werden auch die Spezifitäten gut getrennt.

Für die Simulationen zur allgemeinen Klassifikation ergab sich ein guter Schwellwert ϑ^K bei 1,3. Dieser Wert ist wesentlich für die Klassifikation in (5.98). Wählt man ihn zu gering, ergeben sich zwar weniger (falsch) zugeordnete Konzepte, die nicht im Eingabebild gegeben sind. Es ergeben sich dafür aber auch weniger (richtig) nicht zugordnete Konzepte, die im Eingabebild gegeben sind. Erhöht man den Schwellwert ergeben sich weniger (falsch) nicht zugeordnete Konzepte, die im Eingabebild gegeben sind. Es ergeben sich aber auch weniger (richtig) nicht zugeordnete Konzepte, die nicht im Eingabebild gegeben sind.

Aufgrund von begrenzten Hardwareressourcen werden nur 4 Schichtengruppen eingeführt. Einzelversuche mit reduzierten Mustersätzen zeigen, dass die Spezifität über die Lerniterationen ansteigt, es dann aber auch zu Verpixelungseffekten auf höheren Schichten kommt, und so die Aktivierungsverhältnisse wieder sinken. Verpixelung beschreibt eine kaum noch den Strukturen des Eingabebildes zuzuordnende Aktivierungsverteilung, die bei der Auswertung nur Ergebnisse nahe des statistischen Rauschens hervor-

bringen

6.2.2 Auswertungsdiagramme

Für die Darstellung der Simulationsergebnisse werden verschiedene Diagramme verwendet, die im Folgenden beschrieben werden.

Konzeptaktivierung

Das erste dargestellte Diagrammpaar zeigt die mittlere Konzeptaktivierung c^{CA}:

$$c^{CA}(o, Q) = \frac{1}{|Q|\,|p|} \sum_{q \in Q} \sum_{p} c(q, l(\Omega^{top}, o), p). \tag{6.1}$$

Sie wertet in der obersten Schichtengruppe nur die Schicht aus, die dem zugeordneten Konzept entspricht: $l(\Omega^{top}, o)$. Die gemittelte Aktivität sollte sich in Richtung höherer Schichtengruppen steigern, da die Schichten durch die Lernverfahren spezifischer auf die zugeordneten Eingangsbilder reagieren sollen. In der Praxis tun sie das oft nur in den ersten Lerniterationen/Schichtengruppen, weil die höheren Schichtengruppen spezifisch auf komplexere Inhalte werden, die dann aber seltener vorkommen.

Wie geeignet diese Aktivierungen für eine Klassifikation sind, lässt sich anhand dieser Diagramme noch nicht sagen, da nicht berücksichtigt wird, wie stark die übrigen Schichten aktiviert werden, die dem Konzept nicht zugeordnet sind.

Konzeptaktivierungsverhältnis

Das Konzeptaktivierungsverhältnis c^{CR} beschreibt die Aktivierung von der Schicht, die dem zu prüfenden Konzept zugeordnet ist, bezogen auf die gemittelten Aktivierungen der Schichten, die dem Konzept nicht zugeordnet sind:

$$c^{CR}(o, Q) = \frac{c^{CA}(o, Q)}{\frac{1}{|Q|-1} \sum_{o' \in O(Q) | o' \neq o} c^{CA}(o', Q)}. \tag{6.2}$$

Für eine sichere Klassifikation ist es also wichtig für alle Konzepte, Werte über 1 zu erreichen.

Gesamtaktivierungsverhältnis

Das Gesamtaktivierungsverhältnis c^Q fasst die Konzeptaktivierungsverhältnisse zusammen:

$$c^Q(O, Q) = \frac{1}{|O|} \sum_{o \in O} c^{CR}(o, Q). \tag{6.3}$$

Auch dieser Wert sollte für eine sichere Klassifikation über 1 liegen. Diese Bedingung ist allerdings nur notwendig und nicht hinreichend, da einzelne Konzepte immer noch falsch detektiert werden können.

Exklusive Klassifikation

Für diesen Diagrammtyp wird für einen Mustersatz die exklusive Klassifikation nach Gleichung (5.96) berechnet. Danach wird gezählt, wieviele Muster richtig klassifiziert wurden und dann ins Verhältnis zur Anzahl der Muster im Mustersatz gestellt. Der Wert 1 steht also für eine komplett korrekte Klassifikation und der Wert 0 steht für eine komplett falsche Klassifikation. Der Wert, der statistisches Rauschen beschreibt, liegt dazwischen und hängt von der Anzahl der gegebenen Konzepte ab. Er berechnet sich aus $1/|O|$ und liegt bei den hier gezeigten Simulationen mit drei Konzepten bei $0,\overline{3}$. Für diese Klassifikationsart sollte der Wert stets über dem Rauschmaß liegen.

Allgemeine Klassifikation

Dieses Diagrammpaar zeigt die Ergebnisse der allgemeinen Klassifikation nach Gleichung (5.98). Die klassifizierte Ergebnismenge O^{det} für ein Muster q kann verschiedene Konzepttypen enthalten:

- detektiertes Konzept, das im Muster gegeben ist (true pos)
- detektiertes Konzept, das nicht im Muster gegeben ist (false pos)
- nicht detektiertes Konzept, das im Muster gegeben ist (false neg)
- nicht detektiertes Konzept, das nicht im Muster gegeben ist (true neg)

Diese Konzepttypen werden für alle Muster q eines Mustersatzes Q aufaddiert.

Zusammen ergeben die kumulierten true pos und false neg Werte stets die Summe der im Mustersatz gegebenen Konzepte. Die Anzeige der beiden Konzepttypen ist daher auf diese Summe normiert. Analog ergeben die

kumulierten true neg und false pos Werte stets die Summe der im Mustersatz nicht gegebenen Konzepte. Die Anzeige dieser beiden Konzepttypen ist entsprechend normiert.

Aus den o.g. Konzepttypen könnten noch die sog. „Precision“ und „Recall“ Werte bestimmt werden. Darauf wird an dieser Stelle jedoch verzichtet, weil die Mustersätze recht klein sind und die Berechnung der Werte dafür anfällig ist.

6.2.3 Identitätstest

Der Identitästest soll sicherstellen, dass für das Provadero-Verfahren Eigenschaften gegeben sind, die für Musterklassifikatoren grundlegend sind. Es sollen genau die Muster wiedererkannt werden, die auch gelernt wurden. Dabei wird noch keine Ausprägungsvarianz eingebracht. D.h. alle Muster eines Konzeptes sind identisch (Abbildung 6.7).

Als Ergebnis müssen die Aktivierungswerte für den Test- und Trainingssatz identisch sein. Ebenfalls die Aktivierungsverhältnisse und das Ergebnis des exklusiven und allgemeinen Klassifikators. Die Diagramme in Abbildung 6.8 und Abbildung 6.9 zeigen, dass der Trainings- und Testmustersatz jeweils die gleichen Werte zeigen.

Das Konzeptaktivierungsverhältnis zeigt, dass für alle Konzepte Werte über 1 erreicht wurden. Die Detektion für das „square“-Muster wird sogar extrem spezifisch. Insgesamt ergibt sich für die Exklusivdetektion ab der zweiten Schichtengruppe keine Falschdetektion mehr. Die allgemeine Klassifikation detektiert in der vierten Schichtengruppe alle gegebenen Konzepte richtig. Dafür gibt es dann auch Falschdetektionen der nicht gegebenen Muster, die in der dritten Schichtgruppe noch komplett richtig detektiert wurden.

6.2.4 Positionstest

Auch der Positionstest soll eine grundlegende Eigenschaft des Provadero-Verfahrens sicherstellen: die Detektion soll positionsinvariant möglich sein. Dazu wurde ein Mustersatz generiert, der für jedes Konzept identische Muster in unterschiedlichen Positionen zeigt (Abbildung 6.10). Es wird sichergestellt, dass sich keine Position wiederholt.

Die Ergebnisse sollten für den Trainings- und Testsatz wieder in etwa gleich sein, was in den Abbildungen 6.11 und 6.12 auch nachvollziehbar ist.

Dieser Test zeigt, welchen Einfluss ein veränderter Bildrand hat. Dieser wirkt sich auf die Diffusionsergebnisse und damit auf die Folgeassoziationen aus. Wenn die Bildfläche größer gewählt wird, lässt dieser Effekt wieder

nach. Bei diesen kleinen Mustern spielt der Effekt jedoch eine Rolle. Dieser Effekt ist auch den Ergebnissen der folgenden Simulationen anzurechnen.

Die Aktivierungsverhälnisse liegen deutlich über 1 und die Detektion der exklusiven Klassifikation ist ab der zweiten Schichtengruppe komplett richtig. Auch die allgemeine Klassifikation zeigt sehr gute Ergebnisse.

6.2.5 Rotationstest

Bei diesem Test genügt es nicht mehr, identische Muster in der Bildebene zu finden. Die Muster sind unterschiedlich rotiert (Abbildung 6.13) und so muss das Provadero-Verfahren zeigen, dass sinnvoll generalisiert werden kann.

In Abbildung 6.14 ist zu sehen, dass die Aktivierungsverhältnisse des Testsatzes erstmals hinter denen des Trainingssatzes zurück bleiben. Das ist für Lernverfahren auch normal, wenn auf unbekannte Muster generalisiert werden muss. In Abbildung 6.15 ist zu sehen, dass die Gesamtaktivierung des Trainingssatzes ab der dritten Schichtengruppe schon langsam wieder zurückgeht; während die Detektion des Testsatzes bei der exklusiven Klassifikation sich erst in diesem Bereich deutlich aus dem statistischen Rauschen erhebt. Die allgemeine Klassifikation erreicht bis zu 50% korrekt gefundene gegebene Muster und bis zu 80% korrekt gefundene nicht gegebene Muster.

Für diesen Mustersatz wäre der Einsatz der Rücktransformationsfunktion sinnvoll. Ohne diese Funktion fehlt es dem Provadero-Verfahren an einer Möglichkeit, die Objekte invariant in einer standardisierten Form zu erkennen. Die Erkennung funktioniert also nur über die Generalisierungsfähigkeit des Verfahrens, die es ermöglichen soll, abstrakte Varianzen aufzufangen. Aufgrund des kleinen Trainingssatzes ist es für das Verfahren nicht einfach, die richtigen Eigenschaften für eine gute Rotationsgeneralisierung zu finden.

6.2.6 Skalierungstest

Auch bei diesem Test werden Generalisierungsfähigkeiten des Provadero-Verfahrens gefordert. Die Muster sind in unterschiedlichen Skalierungen gegeben, die sich nicht wiederholen (Abbildung 6.16).

In Abbildung 6.17 ist zu sehen, dass die Aktivierungen des Testsatzes deutlich hinter denen des Trainingssatzes zurück bleiben. Das gilt auch für fast alle Aktivierungsverhältnisse. Die Ergebnisse der Klassifikationen in den Abbildungen 6.17 und 6.18 sind vergleichbar mit dem Rotationstest.

6.2.7 Varianzentest

Für den Varianzentest wurden erstmals für die Muster auch kleine Ausprägungsvarianzen eingeführt und mit zuvor getesteten affinen Varianzen gemischt (Abbildung 6.19).

Die Aktivierungen und Aktivierungsverhältnisse in Abbildung 6.20 zeigen, wie für den Testsatz erst wieder einige Lerniterationen benötigt werden, bis sich die Werte verbessern. In Abbildung 6.21 ist zu sehen, wie für den Testsatz bei der exklusiven Klassifikation erst in der vierten Iteration das Rauschniveau erreicht wird. Die Kurve „hängt“ vorher darunter und ist deshalb zwar statistisch signifikant - allerdings im negativen Sinn. Weitere Simulationen zeigen, dass die Kurve dann schon noch aus dem Rauschbereich herauskommt. Es wäre interessant, die Entwicklung für weitere Schichtengruppen anzusehen. So findet auch die allgemeine Klassifikation in diesem Iterationsbereich keines der gegebenen Muster richtig. Erst in der vierten Iteration werden 18% dieser Muster korrekt erkannt. Die richtig erkannten nicht gegebenen Muster liegen immerhin zwischen 60% und 80%.

Der Varianzentest würde sicherlich von einer realisierten Rücktransformation profitieren können, da dann nicht Rotationsinvarianz und Ausprägungsvarianz gleichzeitig durch die Detektionselemente aufgefangen werden müsste.

6.2.8 Test auf Bild-Film-Diskurs Datensatz

Ein Teil des Bild-Film-Diskurs Datensatz ist in Abbildung 6.22 und Abbildung 6.23 dargestellt. Er wird in dieser Zusammenstellung für zwei „kleine“ Tests verwendet, die zeigen sollen, wie sich das Verfahren unter unterschiedlich stark ausgeprägten Abstraktionsvarianzen verhält. Es schließt sich ein „großer“ Test an, der mit jeweils 20 Trainings- und 10 Testmustern für jeweils drei Konzepte aus dem BFD-Datensatz durchgeführt wird. In Abbildung 6.24 sind beispielhaft für 6 Muster die im Datensatz gegebenen Masken dargestellt. Make 1 umfasst die handelnden Personen, Maske 2 umfasst die Gesichter der handelnden Personen und Maske 3 umfasst handlungsrelevante Objekte. Für das Konzept Eid ist es die zum Eid erhobene Hand, für das Konzept Handschlag sind es die umfassenden Hände und für das Konzept Pieta ist es die getragene Person. Für die Simulationen wurde für das Lernverfahren die personenumfassende Maske verwendet.

Kleiner BFD-Test 1

Aus der BFD-Bilderdatenbank wurden drei Konzepte ausgewählt, die noch relativ geringe abstrakte Varianzen zeigen. Für die drei Konzepte „Eid“, „Handschlag“ und „Amok“ wurden jeweils sechs Trainingmuster und vier Testmuster ausgewählt (Abbildung 6.22).

In Abbildung 6.25 ist zu sehen, dass sich für alle Konzepte bald ausreichende Aktivierungsverhältnisse ergeben. Abbildung 6.26 zeigt, dass sich für die exklusive Klassifikation schon nach der zweiten Lerniteration Werte über dem Rauschniveau ergeben. Die allgemeine Klassifikation erreicht auch gute Werte. Es werden 33% aller gegebenen Muster erkannt. Nach der dritten Lerniteration ergeben sich sogar keine positiven Fehldetektionen mehr.

Kleiner BFD-Test 2

Gegenüber dem vorigen Test wurde das dritte Konzept „Amok“ gegen das Konzept „Pieta“ eingetauscht. In Abbildung 6.23 ist zu sehen, dass diese eine wesentlich höhere Ausprägungsvarianz zeigt.

Abbildung 6.27 zeigt, dass sich nicht für alle Konzepte ausreichende Aktivierungsverhältnisse ergeben. Dennoch erreicht die exklusive Klassifikation in Abbildung 6.28 Werte deutlich über dem Rauschniveau. Auch hier wäre es interessant, noch weitere Schichten einzuführen. Die allgemeine Klassifikation erreicht gute Werte. Es werden 40% aller gegebenen Muster erkannt und zwischen 70% und 80% aller nicht gegebenen Muster nicht erkannt.

Großer BFD-Test

Der große BFD-Test verwendet die gleichen Konzepte, wie der kleine BFD-Test 2. Allerdings sind nun jeweils 20 Trainingsmuster und 10 Testmuster gegeben.

Auch beim großen BFD-Test ergeben sich nicht für alle Konzepte ausreichende Aktivierungsverhältnisse (Abbildung 6.29). Die exklusive Klassifikation in Abbildung 6.30 liegt wieder deutlich über dem Rauschen, läßt jedoch nach der dritten Lerniteration für den Testsatz wieder nach, während sich der Testsatz weiter verbessert. Für maschinelles Lernen wird dieser Effekt „Overfitting“ genannt. Die Trainingsdaten werden immer besser approximiert - allerdings auch das überlagerte Rauschen. Darunter leidet die Generalisierungsfähigkeit und die Werte für die Testdaten werden dadurch schlechter. In dieser Provadero-Konfiguration muss man also darauf achten, rechtzeitig die Lerniterationen einzustellen.

Für die allgemeine Klassifikation ergeben sich noch bessere Leistungen, als für den kleinen BFD-Test. Beim Testmustersatz werden bis zu 50%

der gegebenen Muster richtig erkannt. Dieser Klassifikator scheint von den größeren Datenmengen zu profitieren.

6.2.9 Bewertung

Die Simulationen zeigen, dass der gewählte Ansatz durchaus leistungsfähig ist und als Klassifikator für Bilder mit Ausprägungsvarianzen eingesetzt werden kann. Die Simulationsergebnisse sind im Detail jedoch noch nicht sehr aussagekräftig, da erst mit relativ kleinen Datensätzen experimentiert werden konnte. Man sieht z.B. bei den Gesamtaktivierungsverhälnissen, dass die Testsätze manchmal ein wenig besser als die Trainingssätze sind. Das dürfte bei großen Datensätzen nicht mehr auftreten.

Ob die Verarbeitung von abstrakten Varianzen funktioniert, ist im Varianzentest (Abschnitt 6.2.7) noch nicht sichtbar, da nicht genügend Schichten verarbeitet wurden und keine Rücktransformation zur Verfügung stand. Die Tests mit dem BFD-Datensatz (Abschnitt 6.2.8) zeigen jedoch klar, dass dies möglich ist - insbesondere, wenn die abstrakten Varianzen nicht zu stark sind. Beim Varianzentest wurden nur kleine Muster verwendet, die wenig Merkmale zum Diskriminieren bieten. Die Realweltbilder im BFD-Test bieten viel mehr Merkmale. Möglicherweise gelingt deshalb dort das Generalisieren besser.

Die exklusive Klassifikation liegt bei einigen Simulationen nur knapp über dem Rauschen. Dagegen funktioniert die allgemeine Klassifikation besser. Auf dem (großen) BFD-Datensatz wurde bereits ein klassischer Klassifikator „Viola-Jones“ (Viola and Jones, 2004) evaluiert (Schäfer, 2009). Dort wurden für gegebene Konzepte folgende Detektionsraten ermittelt: Eid 23%, Handschlag 35% und Pieta 17%. Gemittelt ergibt sich eine Rate von 25%. Das Provadero-Verfahren liegt zum Teil deutlich darüber. Durch das lokal arbeitende Viola-Jones-Verfahren ergaben sich sehr hohe Detektionsraten für nicht im Muster gegebene Konzepte. Auch das ist beim Provadero-Verfahren deutlich besser.

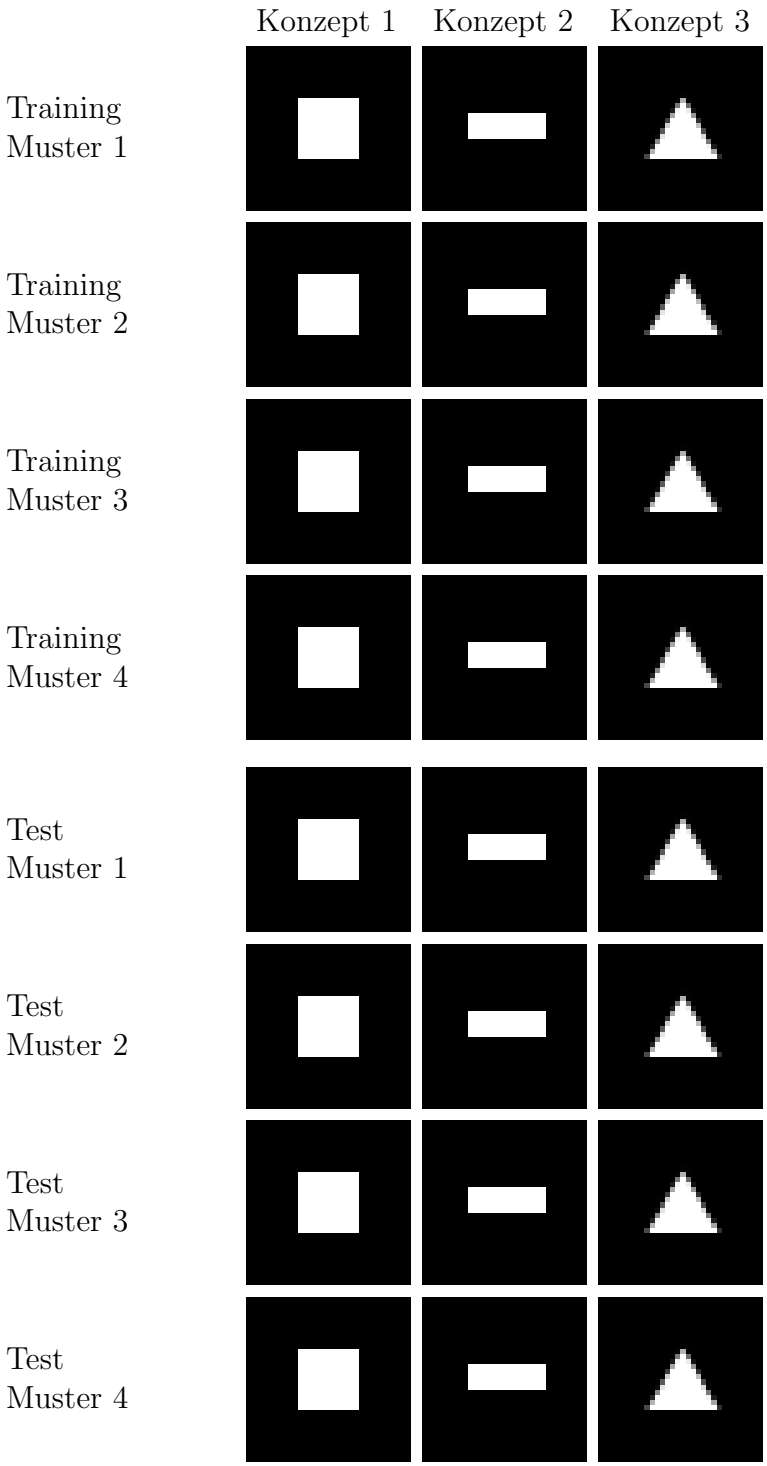


Abbildung 6.7: Muster für Identitätstest (siehe Text)

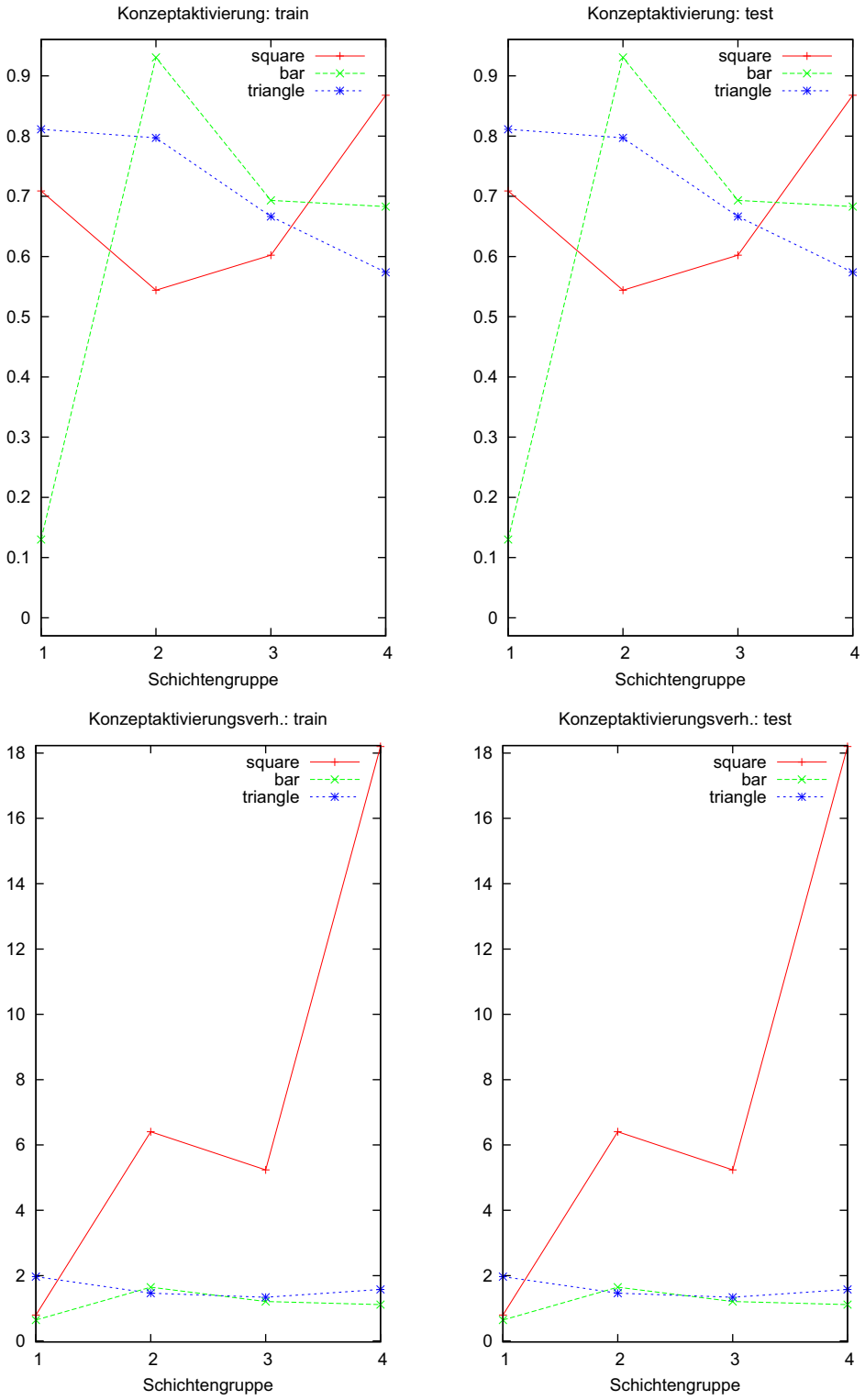


Abbildung 6.8: Ergebnisse (1) für Identitätstest (siehe Text)

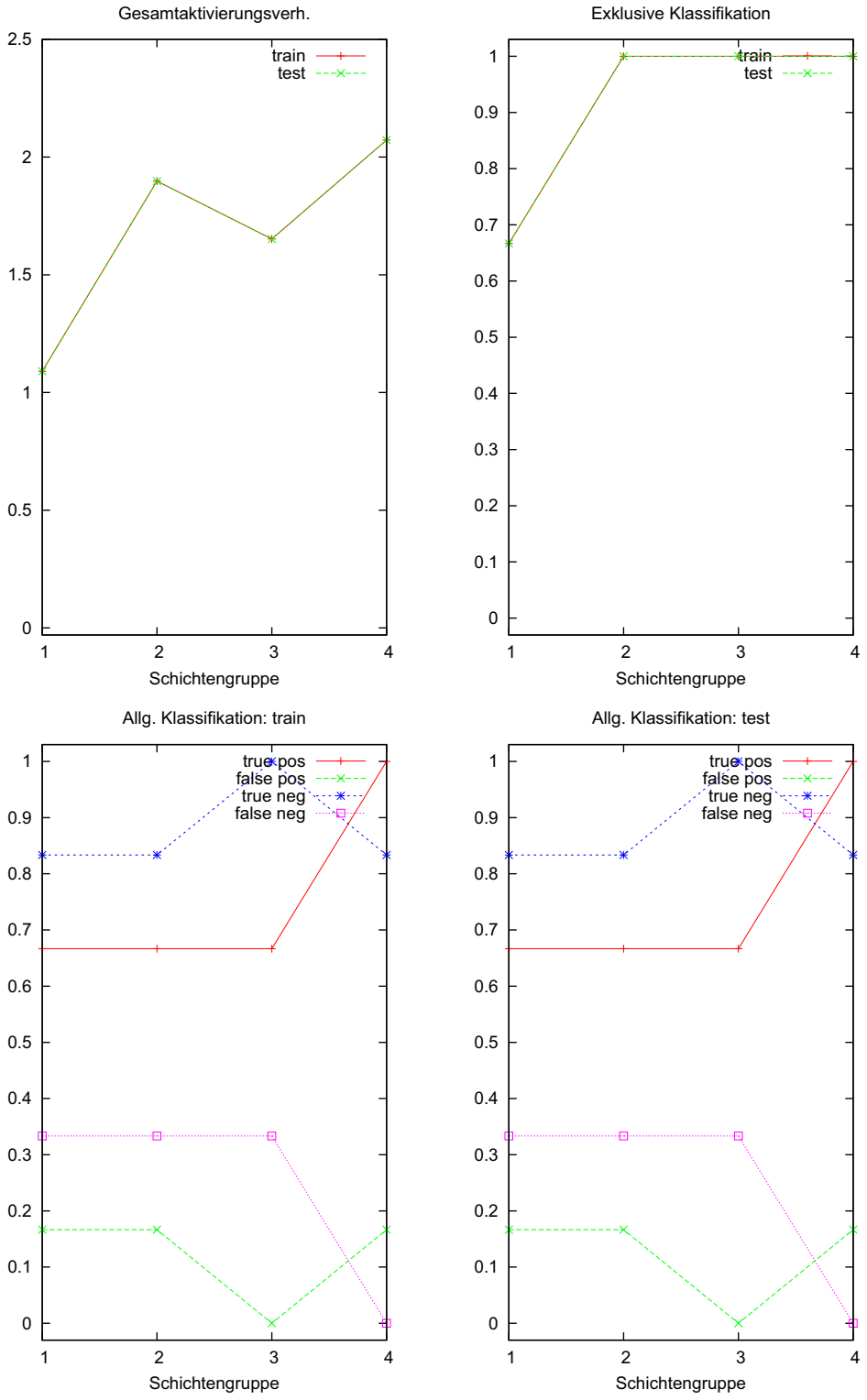


Abbildung 6.9: Ergebnisse (2) für Identitätstest (siehe Text)

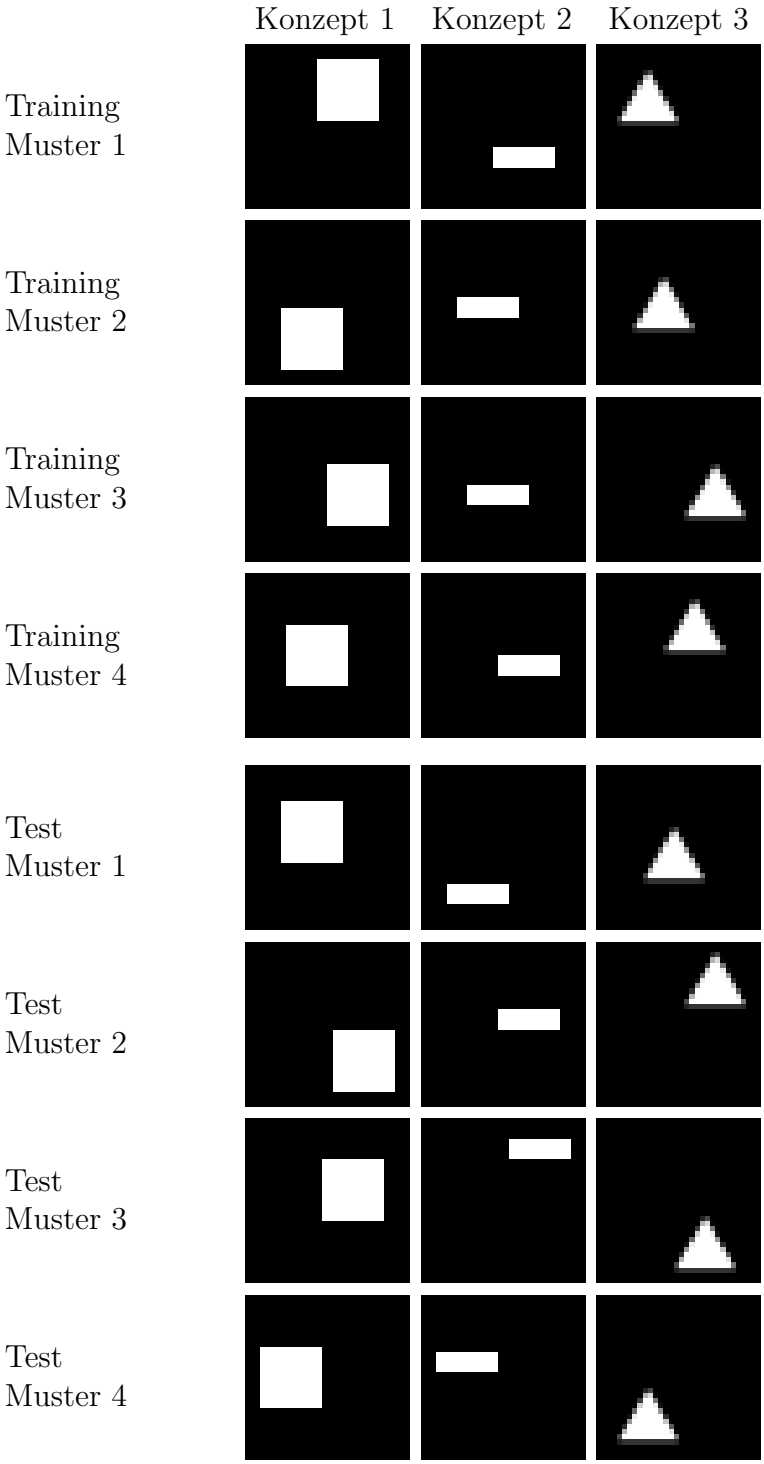


Abbildung 6.10: Muster für Positionstest (siehe Text)

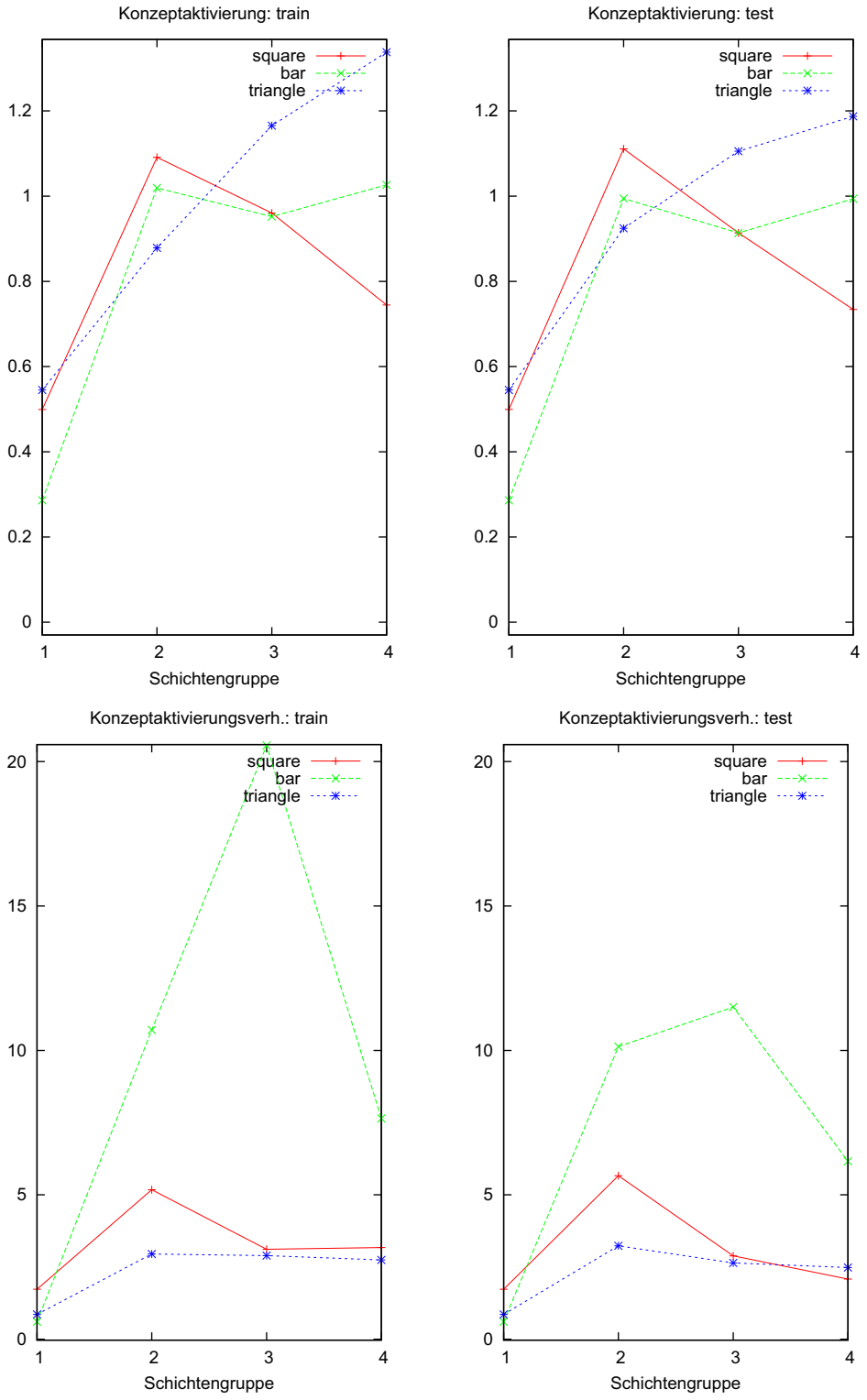


Abbildung 6.11: Ergebnisse (1) für Positionstest (siehe Text)

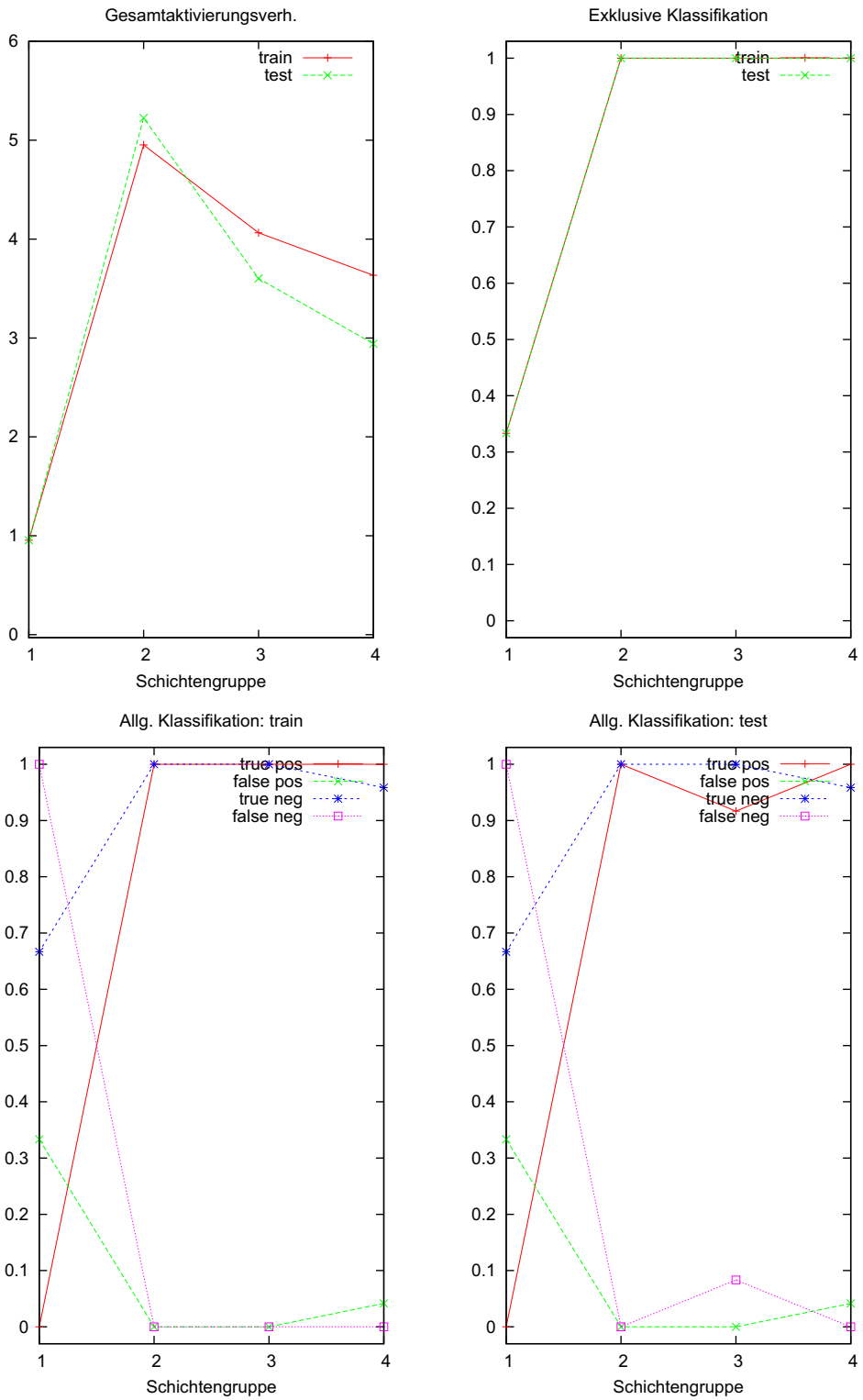


Abbildung 6.12: Ergebnisse 2 für Positionstest (siehe Text)

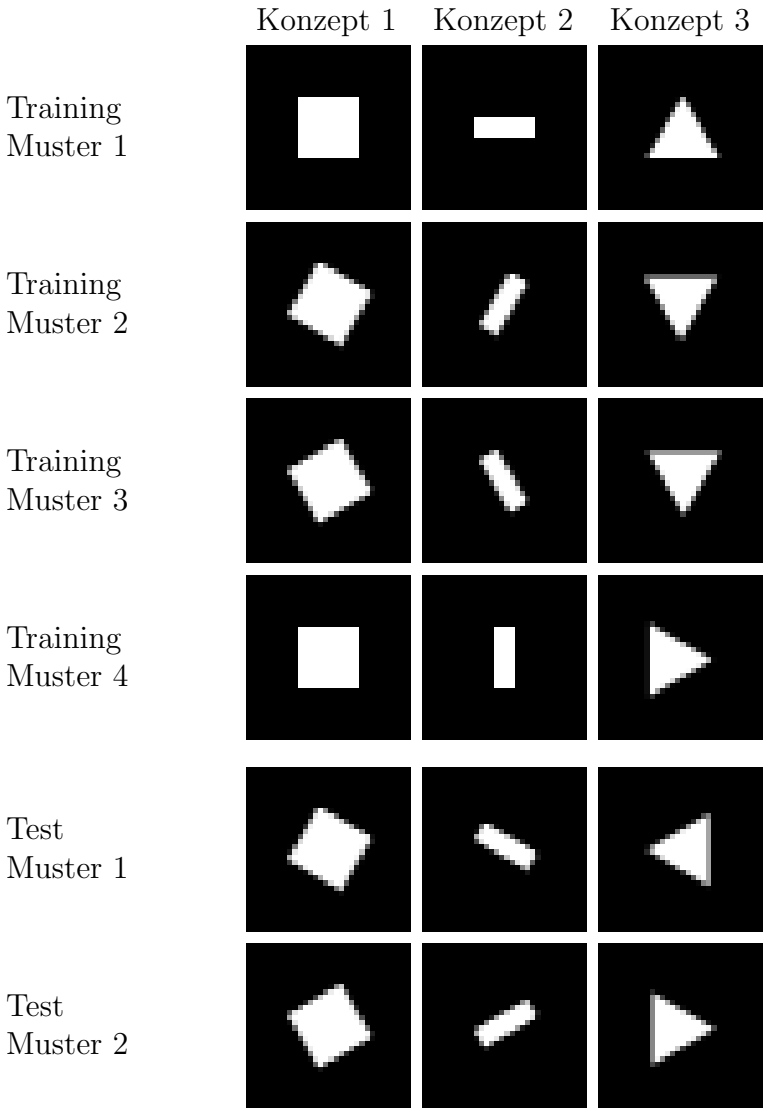


Abbildung 6.13: Muster für Rotationstest (siehe Text)

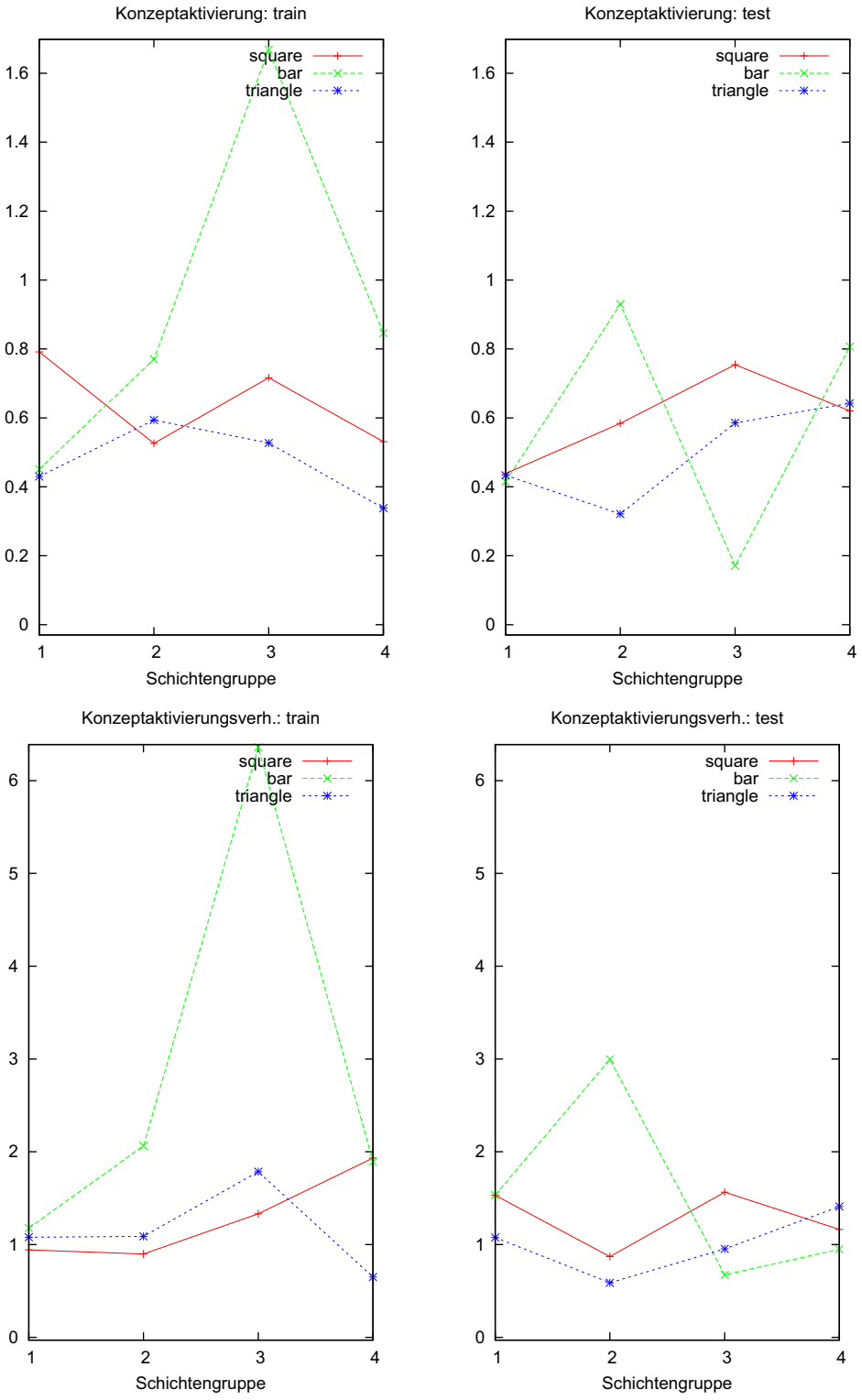


Abbildung 6.14: Ergebnisse (1) für Rotationstest (siehe Text)

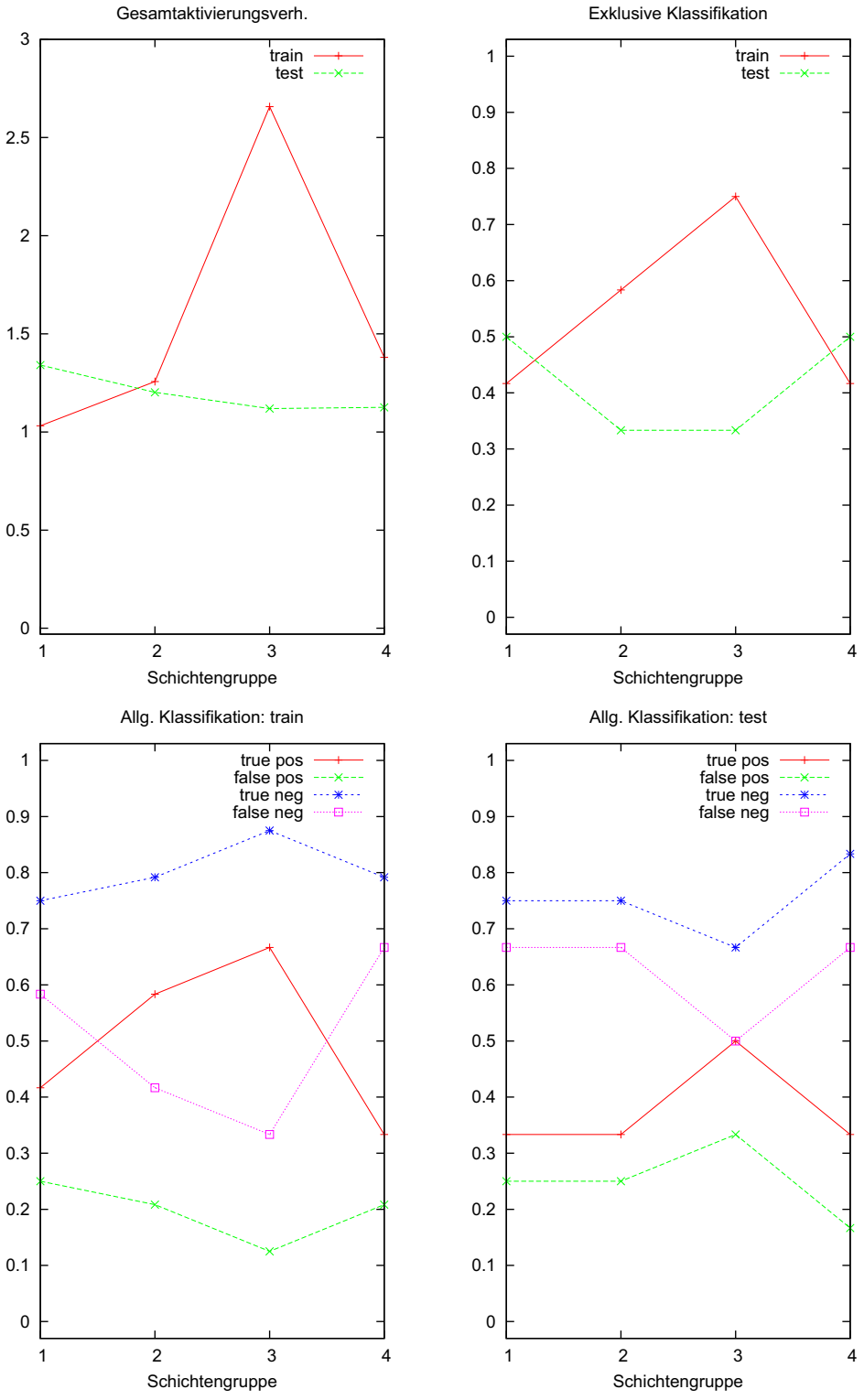


Abbildung 6.15: Ergebnisse (2) für Rotationstest (siehe Text)

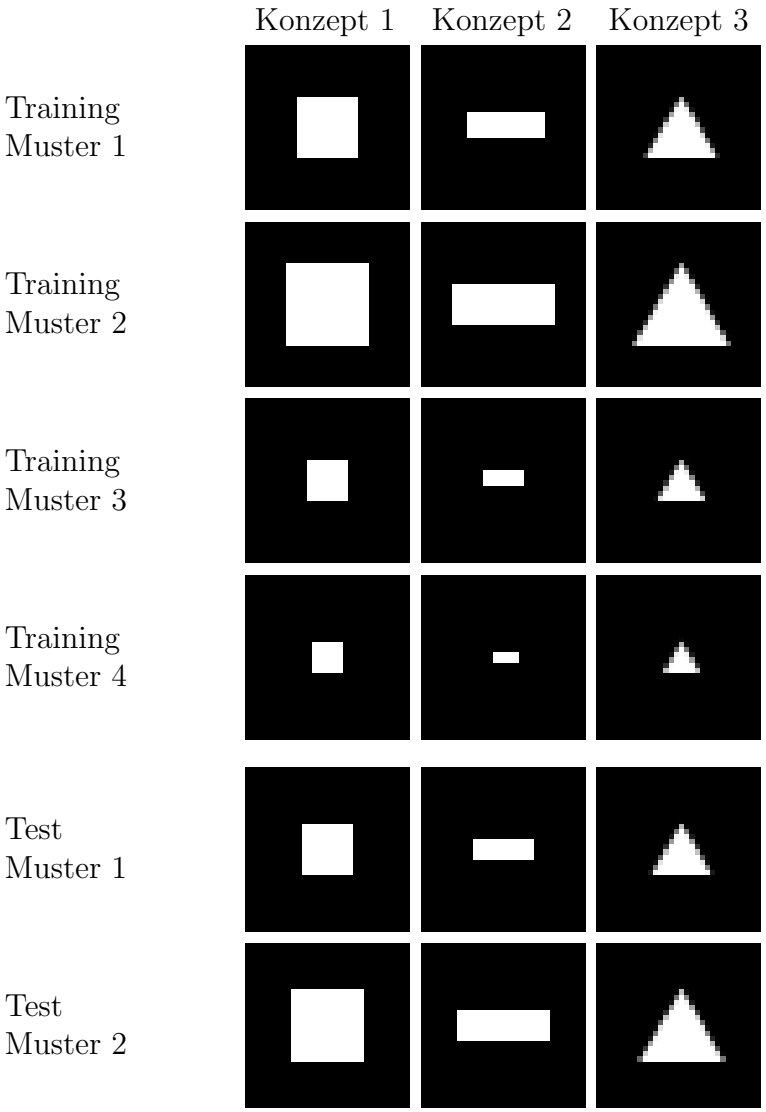


Abbildung 6.16: Muster für Skalierungstest (siehe Text)

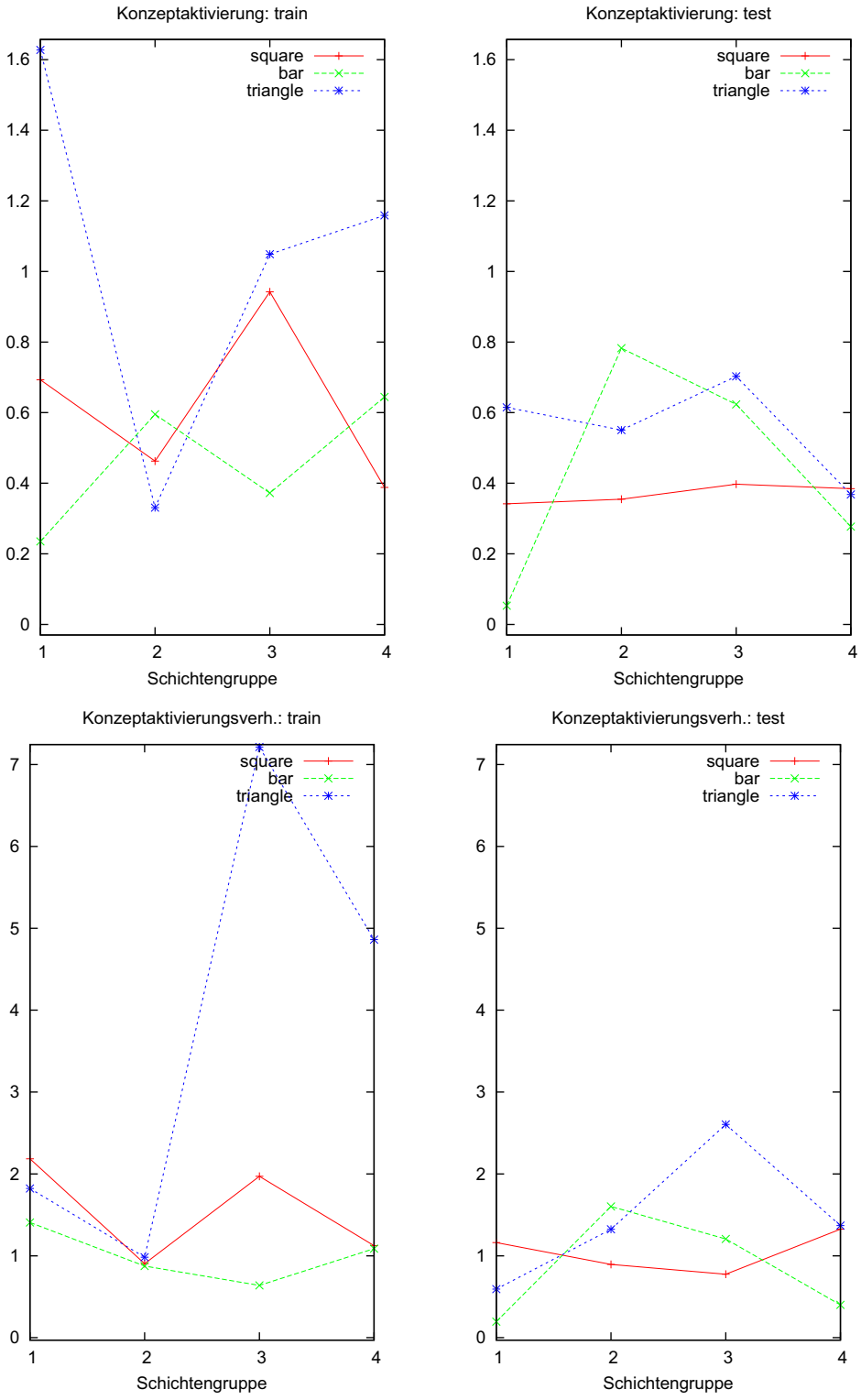


Abbildung 6.17: Ergebnisse (1) für Skalierungstest (siehe Text)

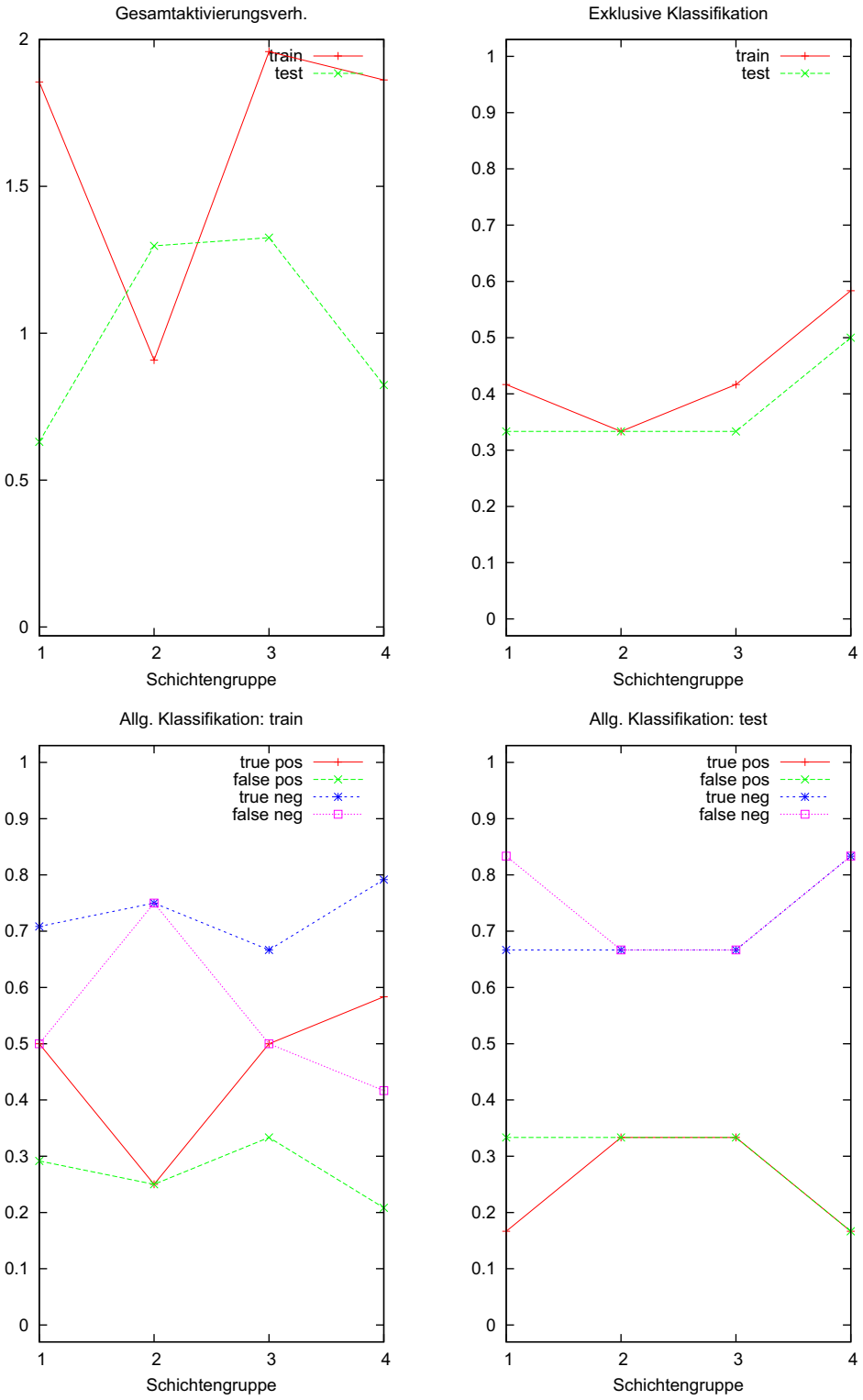


Abbildung 6.18: Ergebnisse (2) für Skalierungstest

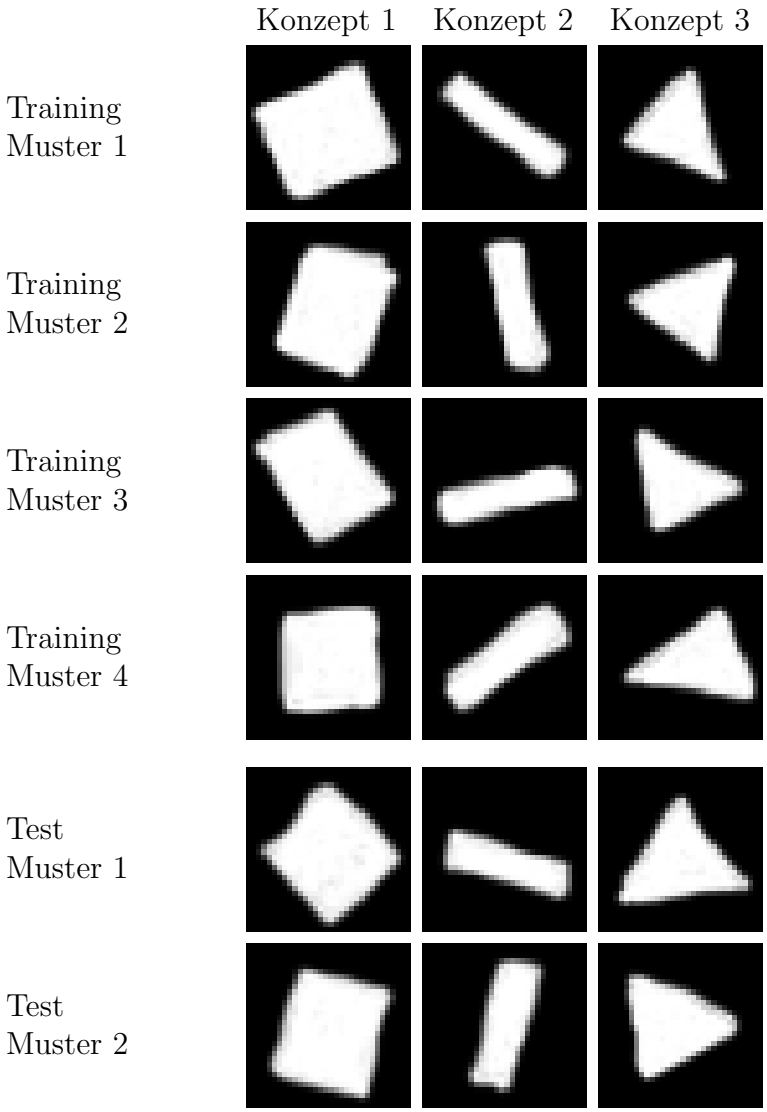


Abbildung 6.19: Muster für Varianzentest (siehe Text)

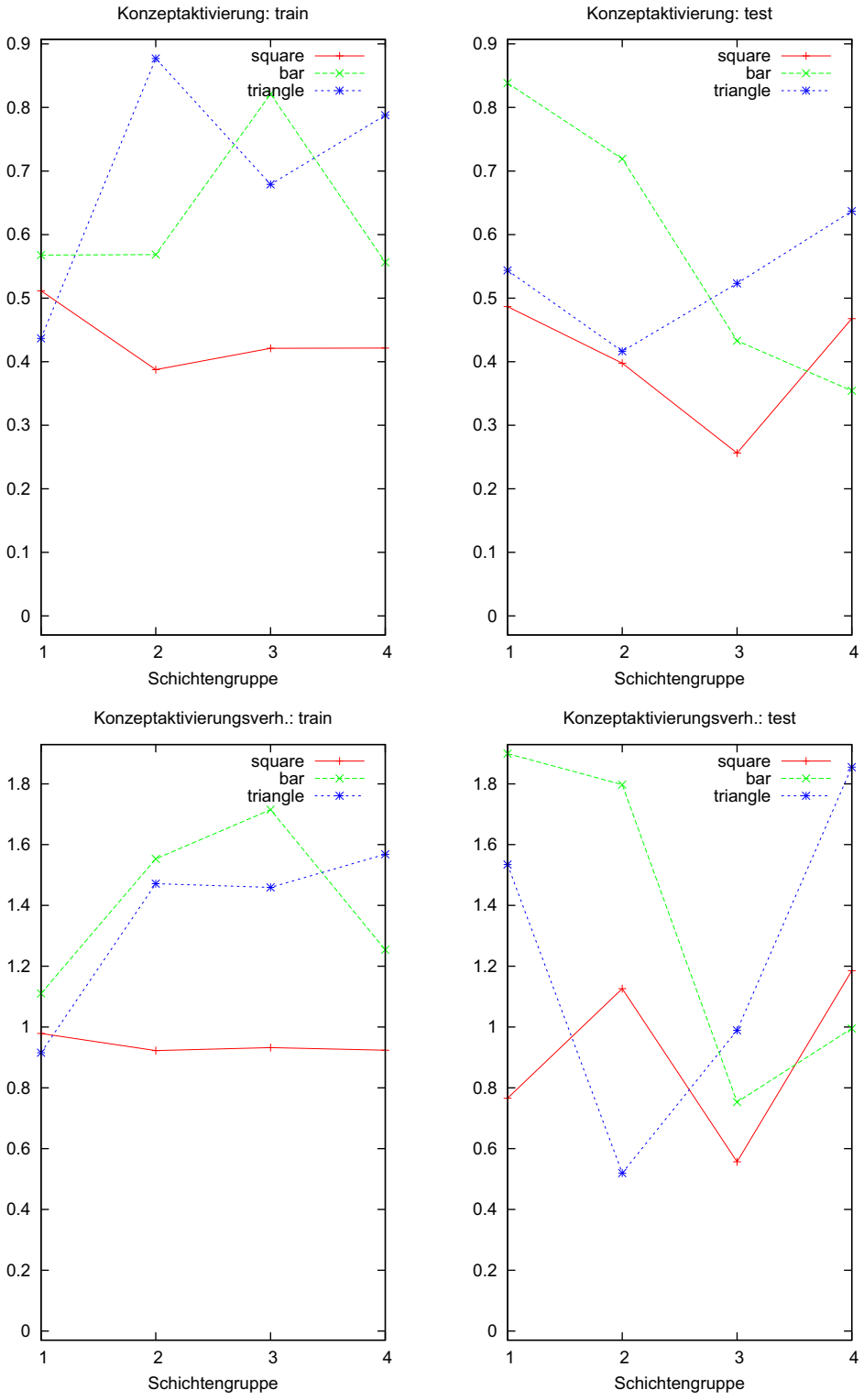


Abbildung 6.20: Ergebnisse (1) für Varianzentest (siehe Text)

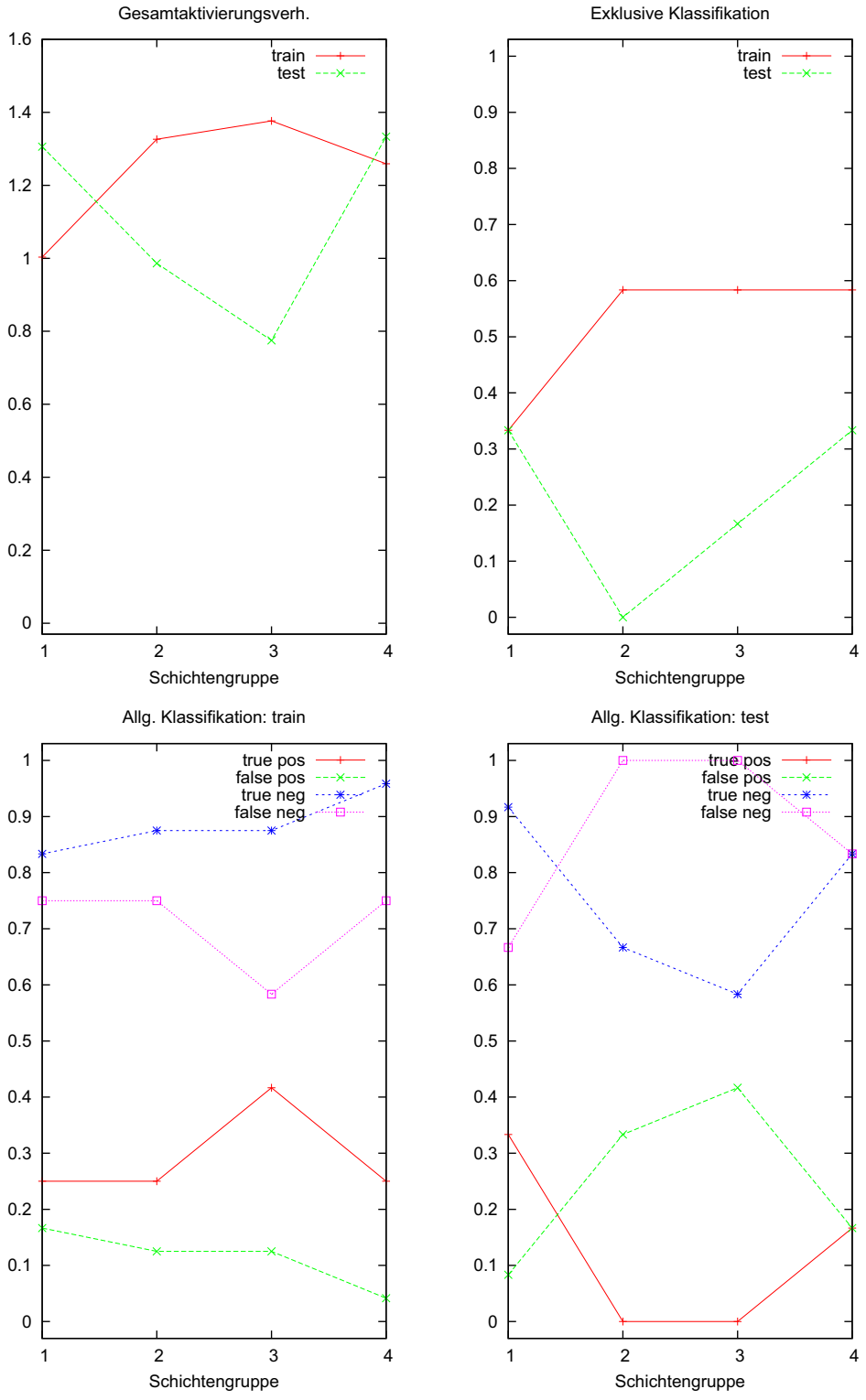


Abbildung 6.21: Ergebnisse (2) für Varianzentest (siehe Text)

Abbildung 6.22: Muster für BFD-Test 1 (siehe Text)

Abbildung 6.23: Muster für BFD-Test 2 (siehe Text)

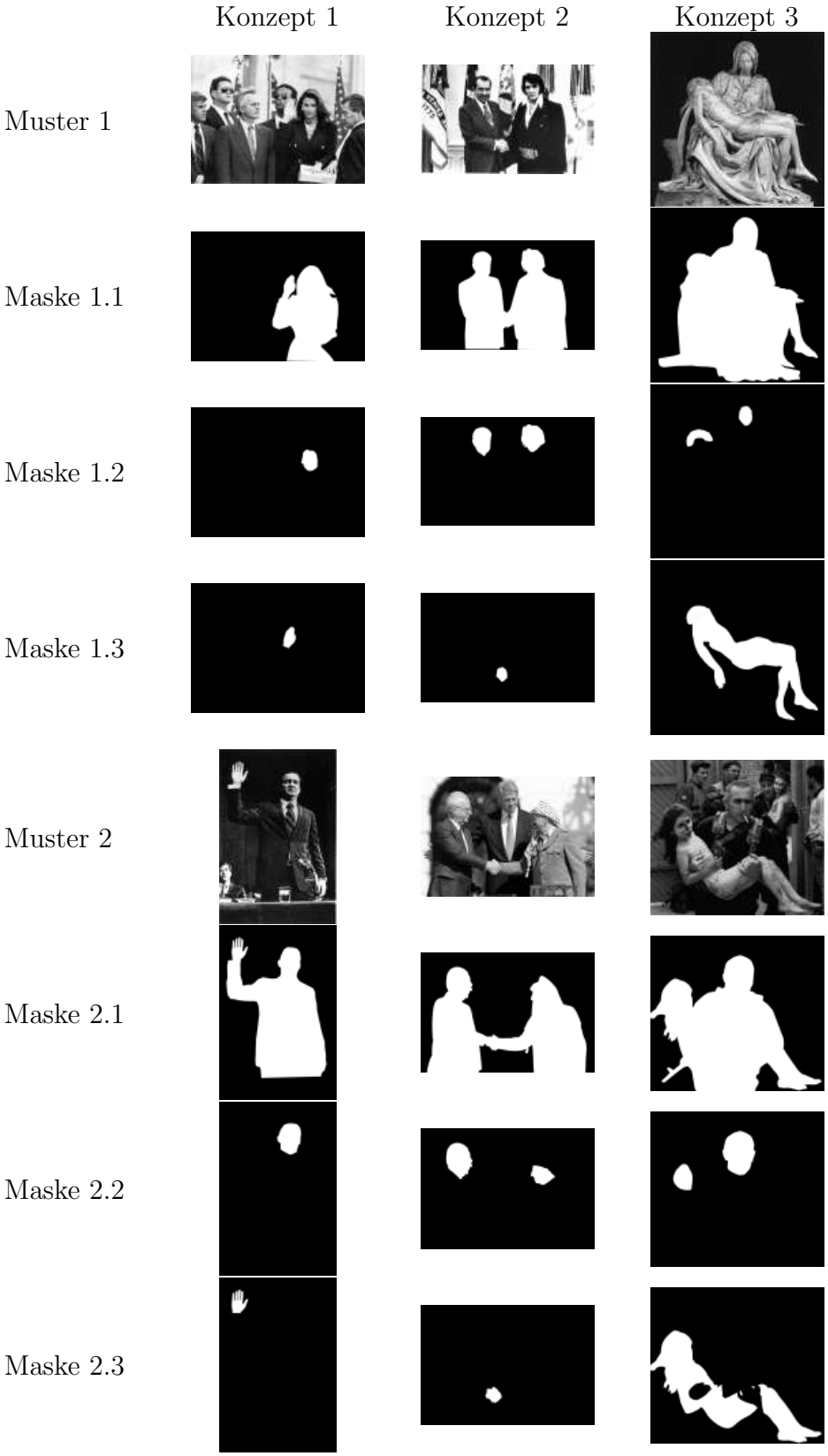


Abbildung 6.24: Verschiedene Masken für den BFD-Test (siehe Text)

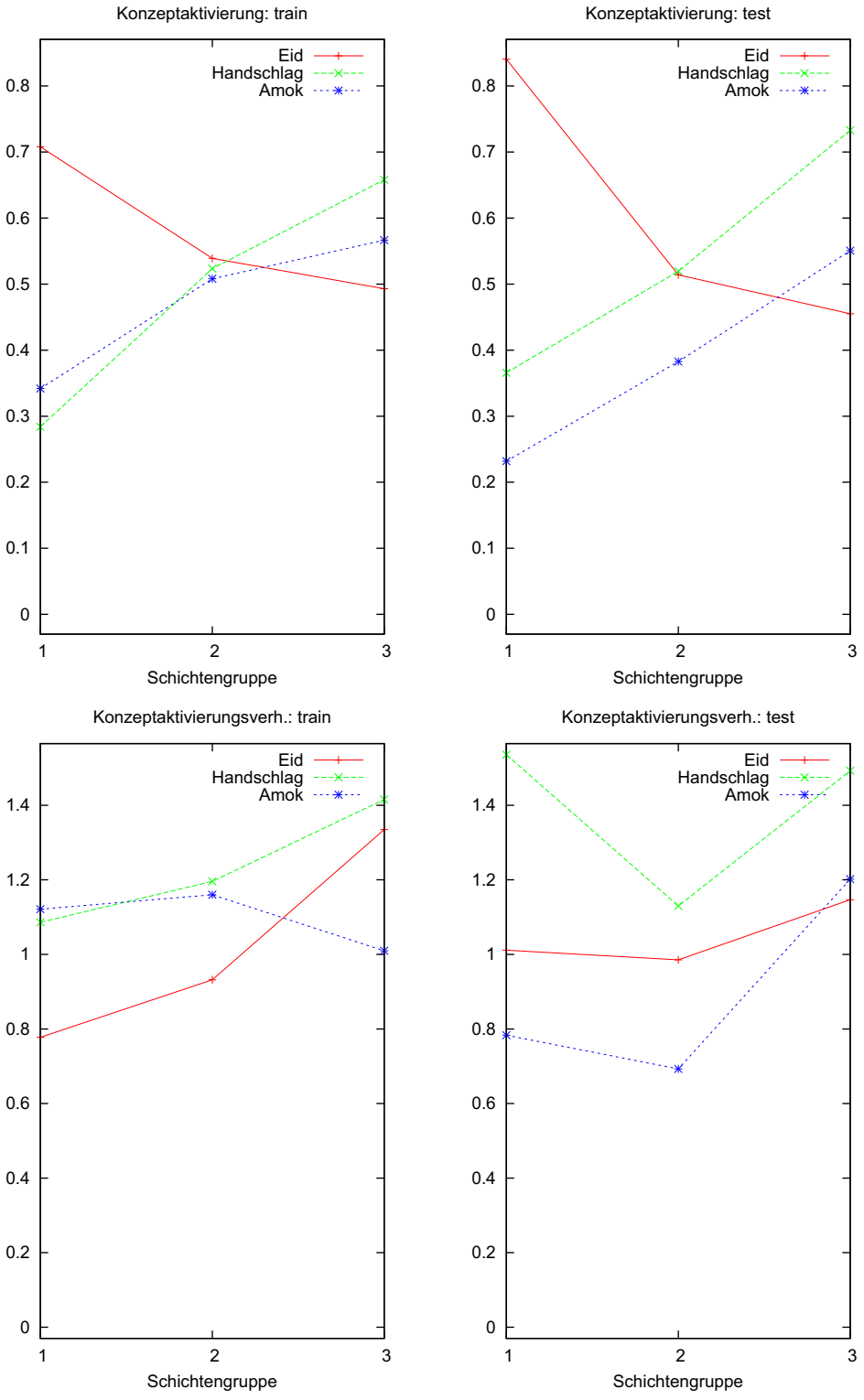


Abbildung 6.25: Ergebnisse (1) Kleiner BFD-Test 1 (siehe Text)

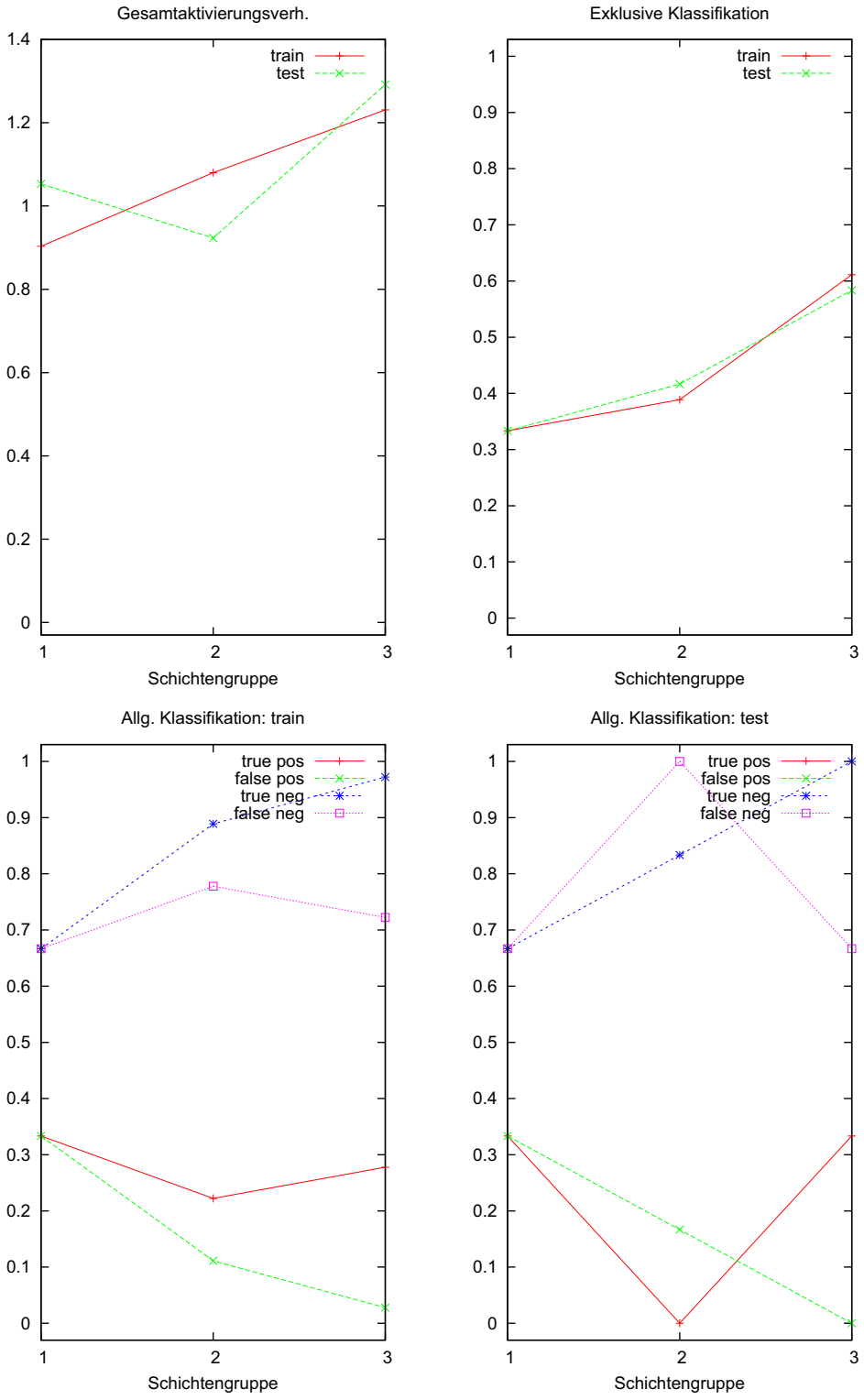


Abbildung 6.26: Ergebnisse (2) Kleiner BFD-Test 1 (siehe Text)

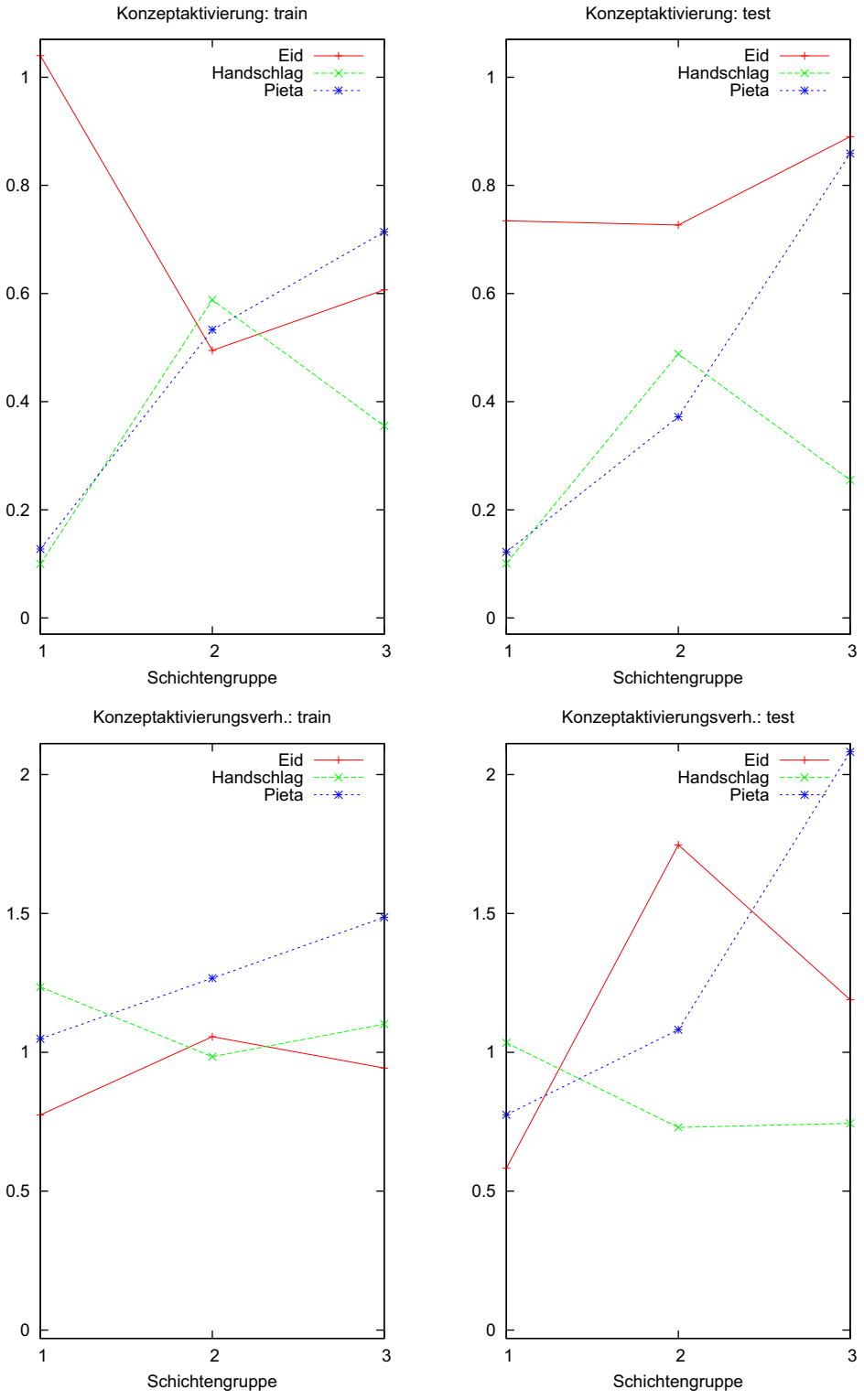


Abbildung 6.27: Ergebnisse (1) Kleiner BFD-Test 2 (siehe Text)

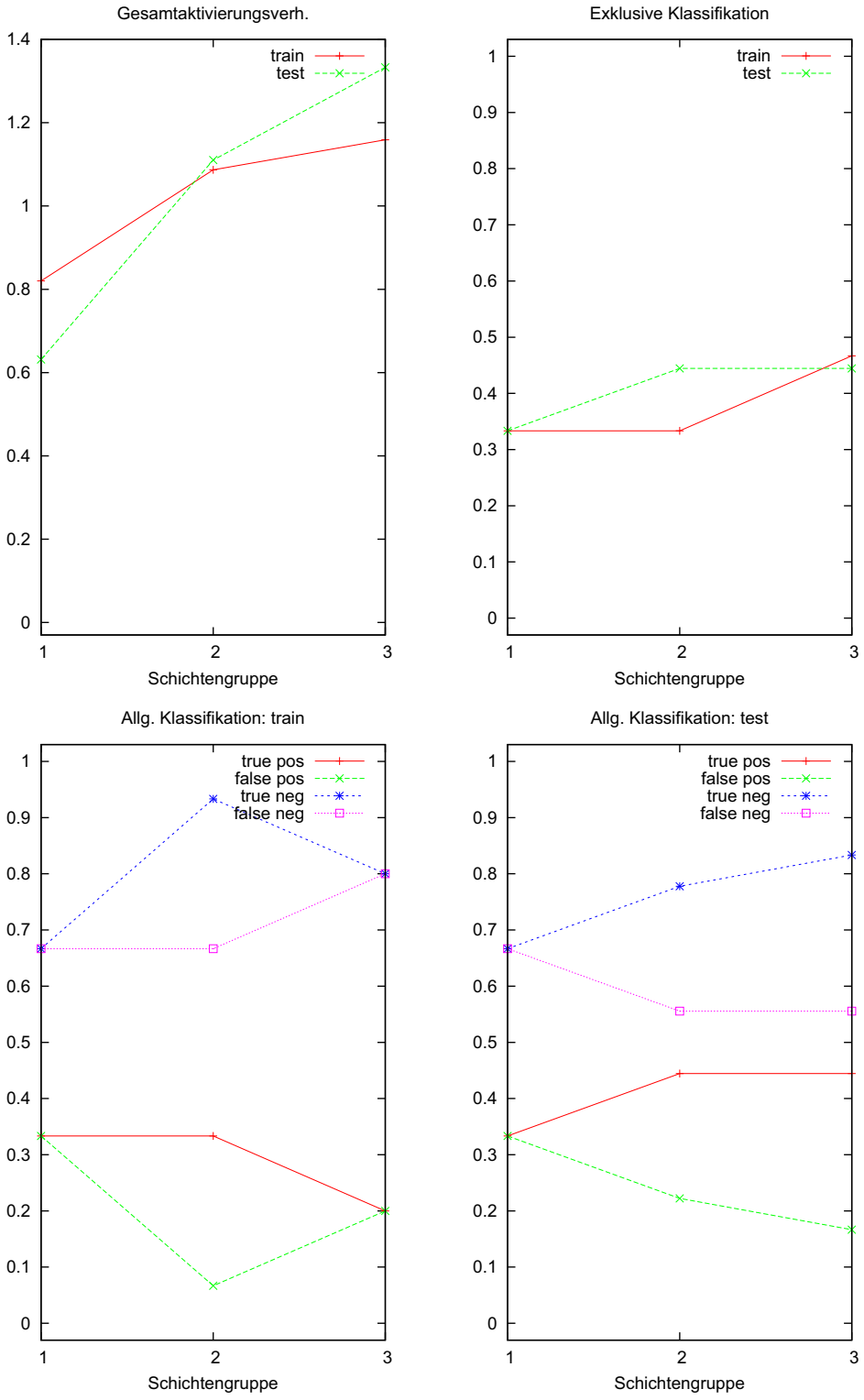


Abbildung 6.28: Ergebnisse (2) Kleiner BFD-Test 2 (siehe Text)

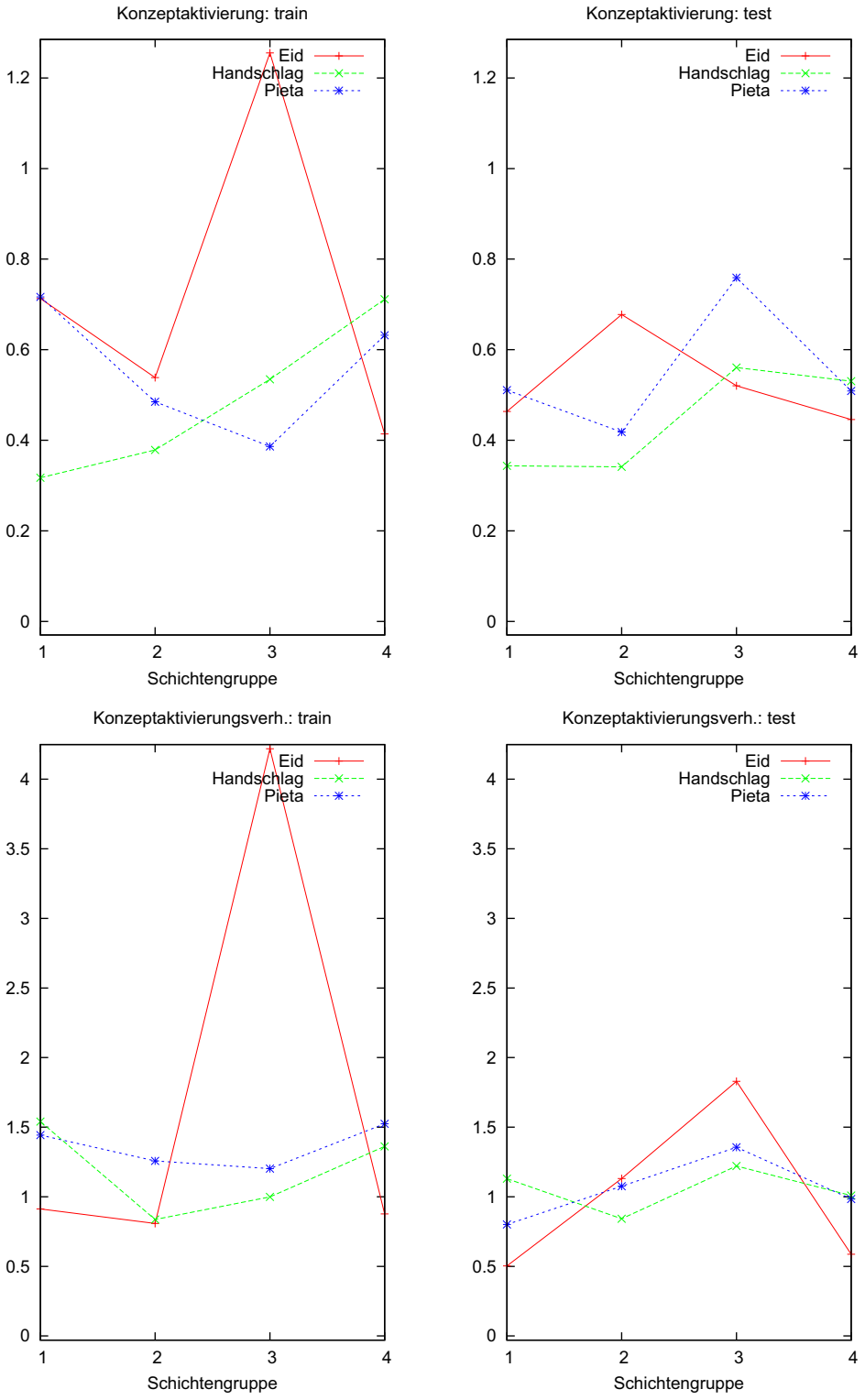


Abbildung 6.29: Ergebnisse (1) Großer BFD-Test (siehe Text)

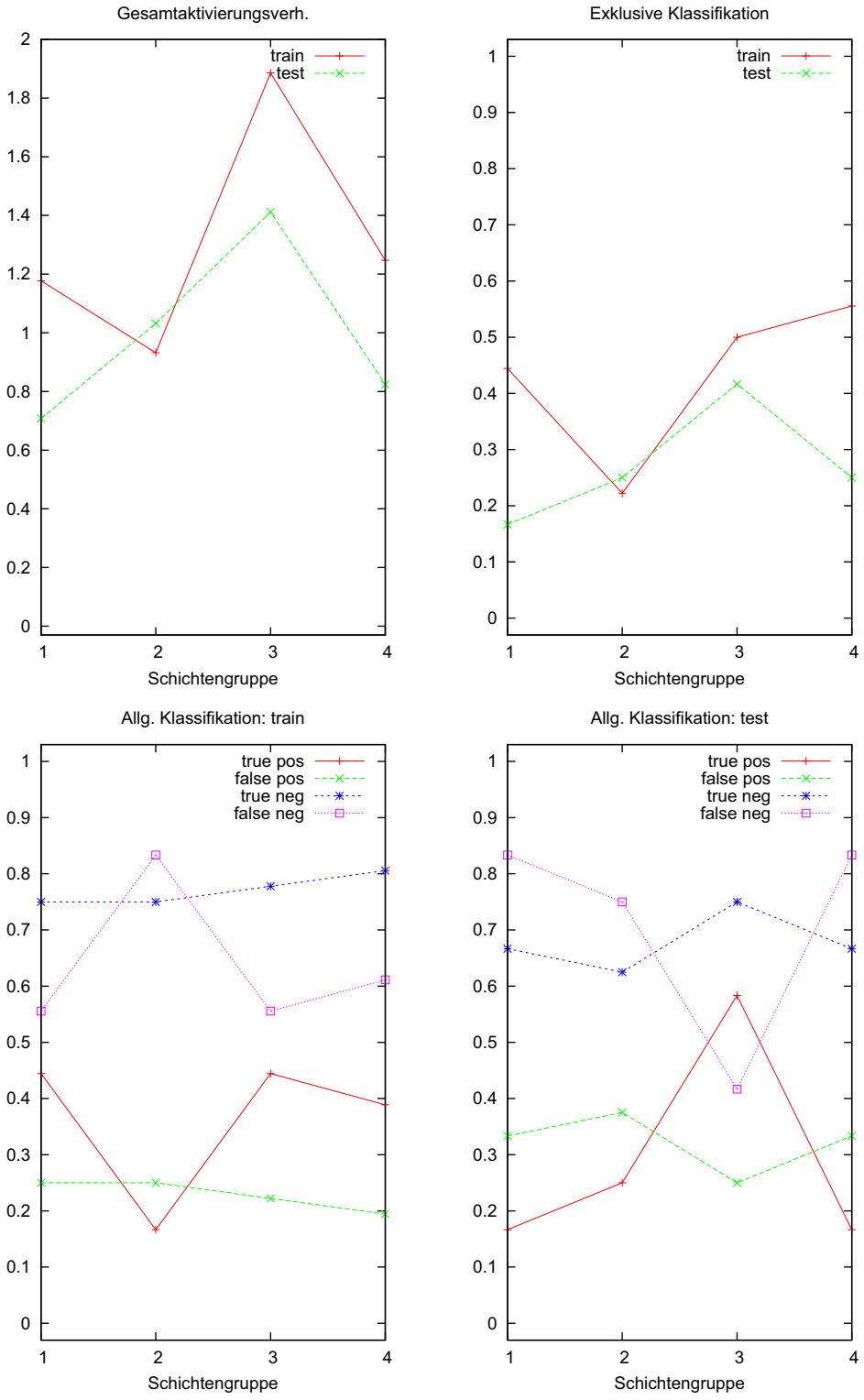


Abbildung 6.30: Ergebnisse (2) Großer BFD-Test (siehe Text)

6.3 Zusammenfassung

Die Simulationen zeigen, dass Provadero eine leitungsfähige Technologie zum Erkennen von Objekten mit Ausprägungsvarianzen bereitstellt. Die Generalisierungsfähigkeiten sind schon ohne die Rücktransformation befriedigend. Bessere Ergebnisse bezüglich Rotation, Skalierung und perspektivischer Verzerrung sind aber erst mit dieser Erweiterung zu erwarten.

Zunächst ist auch noch problematisch, dass das Lernen mit allen Mustern und deren Schichtenstrukturen parallel realisiert ist. Dies führt zu einem sehr hohen Speicherplatzbedarf und das führt wiederum dazu, dass bisher nur auf kleinen Datensätzen simuliert werden kann. Es sollte aber möglich sein, das Lernverfahren so umzubauen, dass mehr sequenzielle Speichernutzung realisiert werden kann. Dann kann mit großen Datenmengen[2] gearbeitet werden, woraus möglicherweise Hinweise abgeleitet werden können, wie Provadero-Komponenten sinnvoll verändert werden können, um noch bessere Ergebnisse zu liefern. Es wäre sicher auch interessant, eine Feldevaluation über Objektabstraktion in der einen Dimension und Objektkomplexität in der anderen Dimension anzustellen.

[2] geeignete Bilderdatenbanken, die Konzepte in unterschiedlichen Ausprägungen darstellen gibt es im Internet, z.B. `http://www.digital-librarian.com/images.html` oder `http://www.cs.cmu.edu/\~{}cil/v-images.html`

Kapitel 7

Zusammenfassung / Bewertung / Ausblick

7.1 Zusammenfassung

Ein funktionierendes universelles Objekterkennungsverfahren brächte für sehr viele Anwendungen einen großen Nutzen. Z.B. gibt es in der Industrie noch sehr viel Automatisierungpotential - überall dort, wo das visuelle System des Menschen einen wesentlichen Beitrag zum Arbeitsprozess liefert.

Bestehende künstliche Vefahren zur visuellen Erkennung, können meist nur mit Objekten mit einer bekannten Ausprägung umgehen. Es besteht dringender Bedarf, diese Fähigkeiten auch auf Objekte mit variierenden Ausprägungen auszuweiten, da viele Realweltobjekte vom Menschen meist Konzepte zugeordnet bekommen, die sich weniger an ihrem Aussehen, als an ihrer Funktion orientieren.

Leider lassen sich die visuellen Systeme der Biologie nicht einfach nachbauen, weil das Zusammenwirken von möglicherweise relevanten Komponenten im Detail sehr komplex ist. Daher soll versucht werden, biologische Verarbeitungsprinzipien herauszuarbeiten und auf dieser Ebene ein Verfahren zu realisieren.

Sowohl für biologische Sehsysteme, als auch für künstliche Verfahren, ergeben sich Herausforderungen. Es wurden wichtige Fragestellungen dieser Herausforderungen identifiziert und in eine Gliederung geformt. Diese umfasst Fragestellungen nach Effizienz, Repräsentation, Integration von Repräsentationen, Lernen und das Verarbeiten von Varianzen.

Diese Fragenstellungen werden im Licht aktueller Forschung am biologischen Vorbild beleuchtet. Die gleichen Fragestellungen werden anschließend, für bestehende künstliche Systeme untersucht.

Die Untersuchung der biologischen Fragestellungen zeigt, dass noch wesentliche Funktionsweisen unklar sind. Es lassen sich jedoch eine Reihe von Verarbeitungsprinzipien finden, die wahrscheinlich eine Rolle spielen.

Die Untersuchung der bestehenden Verfahren zeigt, dass es über lange Zeit keinen geeigneten Ansatz zum Realisieren von Bilderkennung nach biologischen Vorbild gibt. Erst in den letzten Jahren lassen sich einige Fortschritte verzeichenen. Besonders erfolgreich sind dabei wahrscheinlichkeitsbasierte Modelle, die sich auf affin invariante Vorverarbeitung stützen. Damit sind jedoch bisher keine biologisch motivierten Repräsentationshierarchien ermöglicht worden. Dies ist jedoch für eine Konzeptabstraktion wahrscheinlich unumgänglich. Um sich mit den wahrscheinlichkeitsbasierten Modellen nicht in architektonischen Sackgassen zu verlieren, ist es daher wichtig, nach neuen Ansätzen ohne diese Einschränkungen zu suchen.

Es wird das Provadero-Verfahren vorgestellt, welches versucht, wesentliche biologische Verarbeitungsprinzipien umzusetzen. Dabei wird auf die o.g. Fragestellungen eingegangen. Es schließt sich eine detaillierte Darstellung der Provadero-Realisierung an. Dabei werden die Module beschrieben, die an einer Assoziation beteiligt sind: Diffusionsverfahren, Freigabe, Rücktransformation, Skalierung und Projektion. Zusätzlich werden Lern- und Analyseverfahren beschrieben.

Die Evaluierung der neu entwickelten Provadero-Module zeigt, dass das Diffusionsverfahren wunschgemäß funktioniert. Auch die übrigen Module wirken sinnvoll zusammen und können sowohl klassische als auch abstrakte Bildvarianzen verarbeiten. Dies zeigen verschiedene Tests auf Binärwert- und Realweltbildern.

7.2 Beantwortung der wissenschaftlichen Fragestellung

In Abschnitt 1.2.4 wurde die wissenschaftliche Fragestellung beschrieben:

„In diesem Beitrag soll ein Weg gefunden werden, wichtige Funktionsprinzipien biologischen Bilderkennens in ein Verfahren zu integrieren. Welche Funktionsprinzipien erscheinen dazu untersuchenswert? Wie können diese realisiert werden, so dass ein Zusammenwirken möglich wird?“

Es wurden aus dem Stand der Forschung wichtige Herausforderungen identifiziert und relevante Verfahren nach diesen Herausforderungen beurteilt.

Für das Provadero-Verfahren wurden einige Funktionsprinzipien zur Realisierung ausgewählt. Insbesondere sind dies eine explizite Repräsentation

und Verarbeitung von Varianzen. Sowie eine Erkennenshierarchie, die aufsteigend abstraktere Repräsentationen erstellt und auf jeder Repräsentationsstufe zugehörige Varianzen auffangen kann. Weiterhin sollen generell problematische Vorverarbeitungsstufen entfallen.

Zur expliziten Repräsentation und Verarbeitung von Varianzen wurden spezielle Elemente entwickelt, die für variante und invariante Signalanteile getrennte Signalausgänge haben. Es wurden verschieden Module definiert, die eine Verarbeitung dieser getrennten Signale ermöglichen.

Zur hierarchischen Verarbeitung wurden Schichtengruppen eingeführt, die aufsteigend komplexere Kombinationen von Bildstrukturen repräsentieren können. Dadurch dass auf jeder Schicht spezifische Varianzen verarbeitet werden können, ist dies dann auch für Varianzen in der Repräsentationshierarchie möglich.

Nachteile von typischen Vorverarbeitungsstufen sind, dass sie schwierig zu parametrisieren sind, der Raum erkennbarer Eigenschaften schon stark vorgeprägt wird und bezüglich affiner Transfomationen meist schlecht generalisiert werden kann. Es wurde daher ein Verfahren zur Kontextextraktion entwickelt, dass sowohl auf den Daten des Eingabebildes, als auch auf den höheren Schichten eingesetzt werden kann.

Alle Komponenten wurden so entwickelt, dass sie im Provadero-Verfahren zusammenwirken können. Das Verfahren ist in der Lage, Objekte mit abstrakten Varianzen in Bildern zu erkennen. Diese Fähigkeit wird vorab mit Beispielbildern erlernt. Es werden dann explizit variante und invariante Signalanteile repräsentiert. Eine Technik zum Rücktransformieren wurde konzeptionell erstellt. Es wurde gezeigt, dass das Verfahren wichtige grundlegende Eigenschaften wie Robustheit und Generalisierungsverhalten besitzt. Für Bilder mit Realweltinhalten wurde die Einsetzbarkeit demonstriert.

Als Nebenprodukt entstanden mit dieser Arbeit Diffusionsverfahren, die zur visuellen Kontextanalyse eingesetzt werden können. Sie haben gegenüber üblicherweise den eingesetzten Filtermasken den Vorteil, dass sie nicht auf einen starren Bereich beschränkt sind und Bildstruktur explizit beschrieben werden kann. Es gibt auch keine Generalisierungsprobleme, die starre Filtermasken typischerweise haben, wenn z.B. für skaleninvariantes Erkennen Filtermasken in unterschiedlichen Größen eingesetzt werden.

Für das Provadero-Verfahren wurden viele Funktionen neu entwickelt. Die damit erreichten Erkennungsleistungen sind daran gemessen gut. Es ist weitere Entwicklungsarbeit notwendig, um das System in Anwendungen einsetzen zu können.

Wie schon in der Einleitung von Kapitel 4 motiviert wurde, ist es schwierig einzelne Module in einem anderen Kontext zu validieren und optimieren, da es an keiner Stelle typische Schnittstellen gibt, die man in bekannten

Verfahren nutzen könnte. Wenn man beispielsweise die getrennte Repräsentation von varianten und invarianten Signalanteilen in einem hierarchischen Pooling-Verfahren überprüfen wollte, sind spezielle Techniken zu erstellen, die mit starren Filtermasken extrahierte Kontextinfomation an die Elemente weiterreicht. Das macht daraus schon wieder ein spezielles, separat zu untersuchendes Verfahren.

So bleibt die Möglichkeit Varianten des Verfahrens zu untersuchen und daraus Schlüsse auf eine sinnvolle Funktionskombination abzuleiten. Dieser Schritt würde aber den Rahmen dieses Beitrags sprengen.

7.3 Mögliche Provadero-Erweiterungen

Für eine genaue Analyse der Varianzenrepräsentation kann das Provadero-Verfahren so erweitert werden, dass Aktivierungen die zu einer Assoziation beigetragen haben, zurückverfolgt werden können. Eine gute Untersuchungsmöglichkeit, wäre dabei zuzulassen, einzelne Projektionsergebnisse in dieser Hierarchie zu verändern. So könnte die Schar der Muster, auf die ein komplexes Konzept spezifisch reagiert, systematisch „abgefahren“ werden.

Dann sollte für das Verfahren das Rücktransformationsmodul realisiert und evaluiert werden. Nachdem dann ein kompletter Modulsatz zur Verfügung steht, können für einzelne Module neue Realisierungen entwickelt werden und gegenüber den alten evaluiert werden. Dies dürfte deutlich leichter fallen, als das bereits erfolgte Erstellen eines funktionierenden Grundsatzes von Modulrealisierungen. Sie bieten einen guten Ausgangspunkt für weitere Optimierungen.

In einigen Modulen steckt weiteres Optimierungspotenzial. So könnte z.B. das Diffusionsverfahren noch andere Strukturrepräsentationen, als nur den dominanten Umgebungsgradienten und dessen Translation berechnen. Generell wird es eine Herausforderung bleiben, zweidimensionale Informationen auszuwerten, da die in Abschnitt 3.2.1 beschriebenen Probleme von Vorverarbeitungsstufen grundsätzlicher Natur sind und die Natur dafür offenbar auch keine einfach nachzuvollziehende Lösung verwendet (Abschnitt 2.2.3).

Das Freigabemodul ist im Moment noch nicht optimal realisiert. Es werden nur Aktivierungen einzelner Komponenten zur Eignung betrachtet. Dabei wäre es für die Projektion wahrscheinlich viel gewinnbringender, wenn Kombinationen gefunden würden, die so korrelieren, dass sich sinnvolle Spezifitäten ergeben. Dazu ist allerdings eine Korrelationsanalyse über den gesamten Zustandsvektor $\mathbf{x}^{\Omega}$ notwendig. Das wird schon nach wenigen Lerniterationen zu aufwändig. Möglicherweise kann aber ein kombiniertes

Lernverfahren, für Freigabe und Spezifität gefunden werden, was mit ein paar zufällig gewählten Komponenten startet und dann schrittweise Optimierungen vornimmt.

Die Spezifität wird im Moment für jede Schicht separat gelernt. Es ist jedoch denkbar, diese Hauptkomponentenvektoren zu orthogonalisieren, damit sich die zu erkennenden Konzepte besser unterscheiden lassen.

Ein in der Biologie umfassend verwendetes Prinzip ist das der Rückwärtsinhibierung (Abschnitt 2.2.4). Dies wahrscheinlich zur Disambiguierung von mehrdeutigen Informationen verwendete Prinzip könnte auch gut in das Provadero-Verfahren integriert werden. Dazu müssten etablierte Repräsentationen auf ihre vorgeschalteten Repräsentationen zurückwirken können. Dies kann wie mit der Rücktransformation durch „Verformen“ der Strukturbeschreibungen erfolgen - aber auch durch direktes Manipulieren der Ausprägung v und der Konfidenz c. Ein Lernverfahren könnte für gemeinsam aktivierte (Vor-)Repräsentationen eine konzeptspezifische Unterdrückung einstellen.

Abbildungsverzeichnis

Literaturverzeichnis

B. Alexe, T. Deselaers, and V. Ferrari. What is an object? In *Conference Computer Vision and Pattern Recognition (CVPR), 2010 IEEE*, pages 73–80, Comput. Vision Lab., ETH Zurich, Zurich, Switzerland, 2010.

R. Alferez and Y.-G. Wang. Geometric and illumination invariants for object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(6):505–536, 1999.

Y. Amit. *2D Object Detection and Recognition*. MIT Press, 2002.

C. Anderson and D. van Essen. Shifter circuits: a computational strategy for dynamic aspects of visual processing. *Proceedings of the National Academy of Science USA*, 84:6297–6301, 1987.

M. A. Arbib. *The Handbook of Brain Theory and Neural Networks*. A Bradford Book, The MIT Press, Cambridge, Massachusetts, London, England, 1995.

E. Ashbridge, D. I. Perrett, M. W. Oram, and T. Jellema. Effect of image orientation and size on object recognition: Response of single units in the macaque monkey temporal cortex. *Cognitive Neuropsychology*, 17(1/2/3):13–34, 2000.

A. Atiya and P. Baldi. Oscillations and synchronizations in neural networks: An exploration of the labeling hypothesis. *International Journal of Neural Systems*, 1:103–124, 1989.

F. Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3):183–193, 1954.

P. Baldi and R. Meir. Computing with arrays of coupled oscillators: An application to pre-attentive texture discrimination. *Neural Computation*, 2:459–471, 1990.

M. Bar and U. Shimon. Spatial context in recognition. *Perception*, 25:343–352, 1993.

H. B. Barlow. Three points about lateral inhibition. In W. Rosenblith, editor, *Sensory Communication*, pages 782–786. MIT Press, 1961.

H. B. Barlow. Single units and sensation: A neuron doctrine for perceptual psychology. *Perception*, 1:371–394, 1972.

H. B. Barlow. Unsupervised learning. *Neural Computation*, 1:295–311, 1989.

A. Batlle, J. amd Casals, J. Freixenet, and J. Marti. A review on strategies for recognizing natural objects in colour images of outdoor scenes. *Image and Vision Computing*, 18:515–530, 2000.

S. Becker. Learning to categorize objects using temporal coherence. In C. L. Giles, S. J. Hanson, and J. D. Cowan, editors, *Advances in Neural Information Processing Systems*, volume 5, pages 361–368. Morgan Kaufmann, San Mateo, CA, 1993.

A. J. Bell and T. J. Sejnowski. The independent components of natural scenes are edge filters. *Vision Research*, 37:3327–3338, 1997.

R. A. Bergman and A. Adel. *Functional Neuroanatomy*. Mcgraw-Hill Professional, 2005.

P. Berkes and L. Wiskott. Slow feature analysis yields a rich repertoire of complex-cell properties. Cognitive Sciences EPrint Archive (CogPrints) 2804, http://cogprints.ecs.soton.ac.uk/archive/00002804/, Feb. 2003.

H. H. Bernd Jähne. *Computer Vision and Applications*. Academic Press, 2000.

I. Biederman. On the semantics of a glance at a scene. In M. Kubovy and J. Pomerantz, editors, *Perceptual Organization*, Hillsdale, NJ, 1981. Erlbaum.

I. Biederman. Recognition by components: A theory of human image understanding. *Psychological Rev*, pages 115–147, 1987.

I. Biederman. Visual object recognition. In S. Kosslyn and D. Osherson, editors, *Visual Cognition*, pages 121–166, Cambridge, 1995. MIT Press.

I. Biederman. Subordinate-level object classification reexamined. *Psychological Research*, 62:131–153, 1999.

I. Biederman. Recognizing depth-rotated objects: A review of recent research and theory. *Spatial Vision*, 13:241–253, 2001.

I. Biederman and M. Bar. One-shot viewpoint invariance in matching novel objects. *Vision Research*, 39:2885–2899, 1999.

I. Biederman and E. E. Cooper. Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, 23:393–419, 1991.

I. Biederman and E. E. Cooper. Size invariance in visual object priming. *Journal of Experimental Psychology: Human Perception and Performance*, 18:121–133, 1992.

I. Biederman and P. C. Gerhardstein. Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception & Performance*, 19(6):1162–1182, 1993.

I. Biederman, A. L. Glass, and E. W. Stacy. Searching for objects in real-world scenes. *Journal of Experimental Psychology*, 97(1):22–27, 1973.

V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE Transactions on pattern analysis and machine intelligence*, 25(9):1063–1074, 2003.

O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

M. Booth and E. Rolls. View-invariant representations of familiar objects by neurons in the inferior temporal cortex. *Cerebral Cortex*, 8:510–523, 1998.

G. Bouchard and B. Triggs. Hierarchical part-based visual object categorization. In *Conference on Computer Vision and Pattern Recognition*, pages 710–715, 2005.

V. Braitenberg and A. Schüz. *Cortex: Statistics and Geometry of Neural Connectivity*. Springer-Verlag, Berlin, 1998.

C. D. Brody. Slow covariations in neural resting potentials can lead to artefactually fast cross-correlations in their spike trains. *Journal of Neurophysiology*, 80:3345–3351, 1998.

H. Bülthoff and S. Edelman. Psychological support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences USA*, 89:60–64, 1992.

H. H. Bülthoff, S. Edelman, and M. J. Tarr. How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 5:247–260, 1995.

G. Carneiro and D. Lowe. Sparse flexible models of local features. In *Ninth European Conference on Computer Vision*, pages 29–43, 2006.

J. Cerella. Pigeons and perceptrons. *Pattern Recognition*, 19(6):431–438, 1986.

Y. Chen and T. Lin. An investigation into the closure process in human visual perception for line patterns. *Pattern Recognition*, 31(1):1–13, 1998.

Y. Choe. Second order isomorphism: A reinterpretion and its implications in brain and cognitive sciences. In W. D. Gray and C. D. Schunn, editors, *Proceedings 24th Annual Conference of the Cognitive Science Society*, pages 190–195, George Mason University, Fairfax, VA., 2002.

M. A. Cohen and S. Grossberg. Absolute stability of global pattern formation ans parallel memory storage by competitive neural networks. *IEEE Transactions Systems, Man ans Cybernetics*, SMC-13:815–826, 1983.

C. Connor, D. Preddie, J. Gallant, and D. van Essen. Spatial attention effects in macaque area v4. *Journal of Neuroscience*, 17:3201–3214, 1997.

F. Crick and C. Koch. Are we aware of neural activity in primary visual cortex? *Nature*, 375:121–123, 1995.

F. Crick and C. Koch. Consciousness and neuroscience. *Cerebral Cortex*, 8: 97–107, 1998.

G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *International Workshop on statistical Learning in Computer Vision*, 2004.

N. Dalal and Triggs. Histograms of oriented gradients for human detection. *IEEE Computer Society on Computer Vision and Pattern Recognition*, pages 886–893, 2005.

R. Datta, D. Joshi, J. Li, James, and Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys*, 40, 2008.

G. Deco and E. T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44:621–644, 2004.

A. Delorme. Early cortical orientation selectivity: How fast inhibition decodes the order of spike latencies. *Journal of Computational Neuroscience*, 15(3): 357–365, Nov-Dec 2003.

R. Desimone and J. Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18:193–222, 1995.

S. Didas, J. Weickert, and B. Burgeth. Properties of higher order nonlinear diffusion filtering. *Journal of Mathematical Imaging an Vision*, 35:208–226, 2009.

M. Dill and S. Edelman. Translation invariance in object recognition and its relation to other visual transformations. Technical Report A. I. Memo No. 1610, MIT, 1997.

M. Dill and M. Fahle. The role of visual field position in pattern discriminating learning. *Proceedings of the Royal Society of London Series B*, 264:1031–1036, 1997.

R. O. Duda, P. E. Harl, and D. G. Stork. *Pattern Classification.* Wiley John & Sons, 2002.

R. Eckhorn, R. Bauer, W. Jordan, W. Brosch, W. Kruse, M. Munk, and H. J. Reitboek. Coherent oscillations: A mechanism of feature linking in the visual cortex. *Biological Cybernetics*, 60:121–130, 1988.

S. Edelman and H. H. Bülthoff. Orientation dependence in the recognition of familiar and novel views of three-dimensional objects. *Vision Research*, 32 (12):2385–2400, 1992.

S. Edelman and S. Duvdevani-Bar. A model of visual recognition and categorization. *Philosophical Transactions of the Royal Society (Biological Sciences) London*, 352:1191–1202, 1997a.

S. Edelman and S. Duvdevani-Bar. Similarity, connectionism, and the problem of representation in vision. *Neural Computation*, 9(4):701–721, 1997b.

S. Edelman and F. Newell. On the representation of object structure in human vision: evidence from differential priming of shape and location. CSRP 500, University of Sussex., 1998.

S. Edelman, B. Hiles, H. Yang, and N. Intrator. Probabilistic principles in unsupervised learning. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems 14*, Cambridge, MA, 2002.

M. C. M. Elliffe, E. T. Rolls, and S. M. Stringer. Invariant recognition of feature combinations in the visual system. *Biological Cybernetics*, 86:59–71, 2002.

A. K. Engel, P. König, and W. Singer. Direct physiological evidence for scene segmentation by temporal encoding. *Proceedings of the National Academy of Sciences of the USA*, 88:9136–9140, 1991a.

A. K. Engel, A. K. Kreiter, and W. König, P. ans Singer. Synchronization of oscillatory neural responses between striate and extrastriate visual cortical areas of the cat. *Proceedings of the National Academy of Sciences of the USA*, 88:6048–6052, 1991b.

A. K. Engel, P. König, A. K. Kreiter, T. B. Schillen, and W. Singer. Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neurosciences*, 15:218–226, 1992.

M. Everingham, A. Zisserman, C. K. I. Williams, L. V. Gool, M. Allan, C. M. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorkó, S. Duffner, J. Eichhorn, J. D. R. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, T. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, and J. Zhang. *The 2005 PASCAL Visual Object Classes Challenge*. LNAI, Springer, 2006. URL `http://lear.inrialpes.fr/pubs/2006/EZKVAMCDDDDEDFGGJKKL`.

M. Everingham, L. V. Gool, C. K. I. Williams, JohnWinn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88:303–338, 2010.

L. Fei-Fei and P. Perona. A baysian hierarchical model for learning natural scene categories. *IEEE Computer Society Conference on Computer Vision and Patten Recognition*, pages 524–531, 2005.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *Conference on Computer Vision and Pattern Recognition Workshop, pp. 178-178*, pages 178–178, 2004.

J. A. Feldman and D. H. Ballard. Connectionist models and their properties. *Cognitive Science*, 6:205–254, 1982.

D. J. Felleman and D. C. Van Essen. Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, 1:1–47, 1991.

P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Anchorage, 2008.

P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *International Journal of Computer Vision*, 61(1):55–79, 2005.

R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *In CVPR*, pages 264–271, 2003.

R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *International Journal of Computer Vision*, 71(3):273–303, 2007.

V. Ferrari, T. Tuytelaars, and L. Van Gool. Simultaneous object recognition and segmentation by image exploration. In T. Pajdla and J. Matas, editors, *European Conference on Computer Vision (ECCV)*, volume 1, pages 40–54. Springer, May 2004.

S. Fidler, M. Boben, and A. Leonardis. Learning hierarchical compositional representations of object structure. In *Conference on Computer Vision and Pattern Recognition*, 2007.

M. A. Fischler and R. A. Elschlager. The representation and matching of pictorial structures. *IEEE Transactions on Computers*, 22(1):67–92, 1973.

J. Fiser and R. N. Aslin. Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological Science*, 12:499–504, 2001.

P. Földiák. Learning invariance from transformation sequences. *Neural Computation*, 3:194–200, 1991.

P. Földiák and M. Young. Sparse coding in primate cortex. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 895–898. Bradford, 1995.

M. Fukumi, S. Omatu, and Y. Nishikawa. Rotation-invariant neural pattern recognition system estimating a rotation angle. *IEEE Transactions on Neural Networks*, 8(3):568–581, 1997.

K. Fukushima. Cognitron: A self-organizing multilayered neural network. *Biological Cybernetics*, 20:121–136, 1975.

K. Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36:193–202, 1980.

K. Fukushima. Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, 1:119–130, 1988.

K. Fukushima, S. Miyake, and T. Ito. Neocognitron: a neural network model for a mechanism of visual pattern recognition. *IEEE Transactions on Systems, Man and Cybernetics*, pages 826–834, 1983.

I. Gauthier, P. Skudlarski, J. C. Gore, and A. W. Anderson. Expertise for cars and birds recruits brain areas involved in face regognition. *Nature Neuroscience*, 3 (2):191–197, February 2000.

T. Gawne and J. Martin. Response of primate visual cortical v4 neurons to simultaneously presented stimuli. *Journal of Neurophysiology*, 2002.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

P. M. Gochin, G. A. Dorfman, and C. G. Gross. Neural ensemble coding in inferior temporal cortex. *Journal of Neurophysiology*, 71(6):2325–2337, 1994.

J. Gordon and D. G. Lowe. What and where: 3d object recogntion with accurate pose. In J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, editors, *Toward Category-Level Object Recognition*, pages 67–82. Springer, New York, 2006.

K. Grauman and T. Darell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Leraning Research*, 8:725–760, 2007b.

C. M. Gray, P. König, A. K. Engel, and W. Singer. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature (London)*, 338:334–337, 1989.

W. E. L. Grimson. On the recognition of curved objects. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:632–643, 1989.

S. Grossberg. How does a brain build a cognitive code? *Psychological Review*, 87:1–51, 1980.

S. Grossberg. How does the cerebral cortex work? development, learning, attention, and 3-d vision by laminar circuits of visual cortex. *Behavioral and Cognitive Neuroscience Reviews*, 2(1):47–76, 2003.

S. Grossberg and N. P. McLoughlin. Cortical dynamics of three-dimensional surface perception: Binocular and half-occluded scenic images. *Neural Networks*, 10(9):1583–1605, 1997.

S. Grossberg and D. Somers. Synchronized oscillations during cooperative feature linking in a cortical model of visual perception. *Neural Networks*, 4:453–466, 1991.

F. H. Hamker. The role of feedback connections in task-driven visual search. In D. Heinke, G. W. Humphreys, and A. Olson, editors, *Connectionist Models in Cognitive Neuroscience, Proceedings of the 5th Neural Computation and Psychology Workshop (NCPW'98)*, pages 252–261, London, 1999. University of Birmingham England, Springer Verlag.

F. H. Hamker and J. Worcester. Object detection in natural scenes by feedback. In H. e. a. Bülthoff, editor, *Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pages 398–407, Berlin, Heidelberg, 2002. BMCV 2002, LNCS 2525, Springer-Verlag.

S. Harnad. The symbol grounding problem. *Physica D*, 42:335–246, 1990.

R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

R. Hartley and A. Zisserman. *Multiple View Geometry*. MIT Press, Cambridge, MA, 2004.

W. G. Hayward and P. William. Viewpoint dependence and object discriminability. *Psychological Science*, 11(1):7–12, 2000.

D. O. Hebb. *The Organization of Behavior: A Neuropsychological Theory*, chapter Chapter 4, pages 60–78. Wiley, New York, 1949.

D. Heinke and G. W. Humphreys. Attention, spatial representation and visual neglect: Simulating emergent attention and spatial memory in the selective attention for identification model (saim). *Psychological Review*, 110(1):29–87, 2003.

D. Heinke, G. W. Humphreys, and G. diVirgilo. Object localization in 2D images based on kohonen's self-organizing feature maps. *Neurocomputing*, 44:817–822, 2002.

F. Heitger, R. von der Heydt, E. Peterhans, L. Rosenthaler, and O. Kübler. Simulation of neural contour mechanisms: representing anomalous contours. *Image and Vision Computing*, 16:407–421, 1998.

J. B. Hellige and N. Cumberland. Categorical and coordinate spatial processing: More on contributions of the transient/magnocellular visual system. *Brain and Cognition*, 45:155–163, 2001.

G. E. Hinton and Z. Ghahramani. Generative models for discovering sparse distributed representations. *Philosophical Transactions of the Royal Society of London*, B:1177–1190, 1997.

J. Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, 79(8):2554–2558, 1982.

D. Hubel and T. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160:106–154, 1962.

D. Hubel and T. Wiesel. Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat. *Journal of Neurophysiology*, 28: 229–289, 1965.

J. E. Hummel. Where view-based theories break down: The role of structure in shape perception and object recognition. In E. Dietrich and A. Markman, editors, *Cognitive Dynamics: Conceptual Change in Human and Machines*, Hillsdale, NJ, 2000. Erlbaum.

J. E. Hummel and I. Biederman. Dynamic binding in a network for shape recognition. *Psychological Review*, 99:480–517, 1992.

M. Humphrey, G. W. amd Roddoch. Routes to object constancy. implications from neurological impairments of object constancy. *Quarterly Journal of Experimental Psychology*, 36a:385–415, 1984.

G. W. Humphreys, L. Fee, and I. D. Gilchrist. Psychophysical analyses of contour processing in humans: the case for qualitative tests. *Image and Vision Computing*, 16:499–509, 1998.

A. Hyvärinen, J. Hurri, and P. O. Hoyer. *Natural Image Statistics*. Springer Verlag, 2009.

R. A. Jacobs, M. I. Jordan, S. J. Nowlan, and G. E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, pages 79–87, 1991.

P. Jolicoeur. A size-congruency effect in memory for visual shape. *Memory and Cognition*, 15:531–543, 1987.

T. Kadir and M. Brady. Saliency, scale and image description. In *International Journal of Computer Vision*, volume 45(2), pages 83–105. Kluwer Academic Publishers, 2001.

T. Kadir, A. Zisserman, and M. Brady. An affine invariant salient region detector. In *Proceedings of the 8th European Conference on Computer Vision*, Prague, 2004.

E. R. Kandel, J. H. Schwarz, and T. M. Jessell. *Kandel, Eric R. and*. Mcgraw-Hill, 2000.

N. Kanwisher. Neural events and perceptual awareness. *Cognition*, 79:89–113, 2001.

G. Kayaert, I. Biederman, and R. Vogels. Shape tuning in macaque inferior temporal cortex. *Journal of Neuroscience*, 23(7):3016–3027, 2003.

R. Klette, K. Schlüns, and A. Koschan. *Computer Vision: Three-Dimensional Data from Images*. Springer Verlag, Singopore, 1998.

A. Knoblauch and G. Palm. Synchronization of neuronal assemblies in reciprocally connected cortical areas. *Theory in Biosciences*, 122:37–54, 2003.

E. Kobatake and K. Tanaka. Neural selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, 71(3):856–867, 1994.

W. Konen, T. Maurer, and C. von der Malsburg. A fast dynamic link matching algorithm for invariant pattern recognition. *Neural Networks*, 7(6/7):1019–1030, 1994.

P. König and A. K. Engel. Correlated firing in sensory-motor systems. *Current Opinion in Neurobiology*, 5(4):511–519, 1995.

P. König and T. B. Schillen. Stimulus-dependent assembly formation of oscillatory responses: I. synchronisation. *Neural Computation*, 3:155–166, 1991.

B. Kosko. Biderectional associative memories. *IEEE Transactions on Systems, Man and Cybernetics*, 18(1):49–60, Feburary 1988.

S. M. Kosslyn, C. F. Chabris, and O. Marsolek, C. J. Koenig. Categorical versus coordinate spatial relations: computational analysis and computer simulation. *Journal of Experimental Psychology: Human Perception and Performance*, 18: 562–577, 1992.

M. Lades, J. C. Vorbrüggen, J. Buhmann, J. Lange, C. v.d. Malsburg, R. P. Würtz, and W. Konen. Distortion invariant object recognition in the dynamik link architecture. *IEEE Transactions on Computers*, 42(3):300–311, 1993.

Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1:541–551, 1989.

Y. LeCun, O. Matan, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, L. D. Jackel, and H. Baird. Handwritten zip code recognition with multilayer networks. In *Proceedings of the 10th International Conference on Pattern Recognition*, Los Alamitos, CA, 1990. IEEE Computer Science Press.

Y. LeCun, F. J. Huang, and L. Bottou. Learning methods for generic object recognition with invariance to pose and lighting. In *CVPR 2004*. IEEE Press, 2004.

W. B. Levy and R. A. Baxter. Energy efficient neural codes. *Neural Computation*, 8(3):531–544, 1996.

M. S. Lew. Content-based multimedia information retrieval: State of the art and challenges. *ACM Trans. Multimedia Computing Communications and Applications*, 2:1–19, 2006.

Z. Li. A neural model of contour integration in the primary visual cortex. *Neural Computation*, 10:903–940, 1998.

H. Liang and H. Wang. Top-down anticipatory control in prefrontal cortex. *Theory in Biosciences*, 122(1):70–86, 2003.

T. Lindeberg. *Scale-Space Theory in Computer Vision*. Kluwer Academic Publishers, 1994.

R. Linsker. How to generate ordered maps by maximizing the mutual information between input and output signals. *Neural Computation*, 1:402–411, 1989.

R. Linsker. Local synaptic learning rule suffice to maximize mutual information in a linear network. *Neural Computation*, 4(691-702), 1992.

Z. Liu, D. Knill, and D. Kersten. Object classification for human and ideal observers. *Vision Research*, 35(4):549–568, 1995.

N. K. Logothetis. Object vision and visual awareness. *Current Opinions in Neurobiology*, 8:536–544, 1998.

N. K. Logothetis, J. Pauls, H. H. Bülthoff, and T. Poggio. View-dependent object recognition by monkeys. *Current Biology*, 4:401–414, 1994.

N. K. Logothetis, J. Pauls, and T. Poggio. Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, 5:552–563, 1995.

D. R. Lovell, T. Downs, and A. C. Tsoi. An evaluation of the Neocognitron. *IEEE Transactions on Neural Networks*, 8(5):1090–1105, September 1997.

D. Lowe. Object recognition from local scale-invariant features. *Proceedings of the International Conference on Computer Vision*, 2:1150–1157, 1999.

D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

S. J. Luck, L. Chelazzi, S. A. Hillyard, and R. Desimone. Mechanisms of spatial selective attention in areas v1, v2 and v4 of macaque visual cortex. *Journal of Neurophysiology*, 77:24–42, 1997.

J. Lücke. Macrocolumns as decision units. In J. R. Dorronsoro, editor, *International Conference on Artificial Neural Networks (ICANN)*, pages 57–62. Springer, 2002.

W. Maass and C. M. Bishop. *Pulsed Neural Networks.* Bradford Books, 1999.

S. Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way.* Academic Press, 2009.

D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information.* W.H. Freeman & Company, New York, 1982.

D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three dimensional structure. *The Proceedings of the Royal Society, London*, 200:269–294, 1978.

J. McClelland and D. Rumelhart. An interactive activation model of context effects in letter perception. *Psychological Review.*, 88:375–407, 1981.

B. W. Mel. SEEMORE: Combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, 9:777–804, 1997.

B. W. Mel and J. Fiser. Minimizing binding errors using learned conjunctive features. *Neural Computation*, 12:731–762, 2000.

W. H. Merigan and J. H. R. Maunsell. How parallel are the primate visual pathways. *Annual Review of Neuroscience*, 16:369–402, 1993.

K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *Journal of Computer Vision*, 60(1):63–86, 2004.

A. D. Milner and M. A. Goodale. *The Visual Brain in Action*. Psychology Series. Oxford University Press, Oxford, 1995.

A. D. Milner and M. A. Goodale. *The Visual Brain in Action*. Oxford University Press, 2006.

P. M. Milner. A model for visual shape recognition. *Psychological Review*, 81: 521–535, 1974.

Y. Miyashita and H. S. Chang. Neuronal correlate of pictoral short-term memory in the primate temporal cortex. *Nature*, 331:68–70, 1988.

Y. Miyashita, A. Date, and H. Okuno. Configuational encoding of complex visual forms by single neurons of monkey temporal cortex. *Neuropsychologia*, 31(10): 1119–1131, 1993.

Y. Moses, Y. Adini, and S. Ullman. Face recognition: The problem of compensating for illumination changes. In *Proceedings of the European Conference on Computer Vision*, pages 286–297, 1994.

V. B. Mountcastle. The columnar organization of the neocortex. *Brain*, 120: 701–722, 1997.

M. Mozer. *The perception of multiple objects*. MIT Press, Cambridge, MA, 1991.

D. Mumford. On the computational architecture of the neocortex. II The role of corticocortical loops. *Biological Cybernetics*, 66:241–251, 1992.

D. Mumford. Neuronal architectures for pattern-theoretic problems. In C. Koch and J. L. Davis, editors, *Large-Scale Neuronal Theories of the Brain*, pages 125–152. MIT Press, Cambridge, MA, 1994.

H. Nakano and T. Saito. Grouping synchronization in a pulse-coupled network of chaotic spiking oscillators. *IEEE Transactions on Neural Networks*, 15(5): 1018–1026, 2004.

T. A. Nazir and J. K. O'Regan. Some results on translation invariance in the human visual system. *Spatial Vision*, 5:81–100, 1990.

F. N. Newell. Stimulus context and view dependence in object recognition. *Perception*, 27:47–68, 1998.

F. N. Newell and J. M. Findlay. The effect of depth rotation on object identification. *Perception*, 26:1231–1257, 1997.

J. C. Niebles and L. Fei-Fei. A hierarchical model of shape and appearance for human action classification. *Proceedings of IEEE Intern. Conf. in Computer Vision and Pattern Recognition(CVPR).*, 2007.

H. Niemann, editor. *Pattern Analysis and Understanding*. Springer-Verlag Berlin and Heidelberg, 1990.

E. Oja. A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, 15:267–273, 1982.

S. Oka, G. J. van Tonder, and Y. Ejima. A vep study on visual processing of figural geometry. *Vision Research*, 41:3791–3803, 2001.

B. Olshausen, C. Anderson, and D. van Essen. A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *Journal of Neuroscience*, 13(4700-4719), 1993.

B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

B. A. Olshausen, C. H. Anderson, and D. C. van Essen. A multiscale dynamic routing circuit for forming size- and position-invariant object representations. *Journal of Computational Neuroscience*, 2:45–62, 1995.

A. Opelt and A. Zisserman. A boundary-fragment-model for object detection. *Ninth European Conference on Computer Vision*, pages 575–588, 2006.

M. W. Oram and D. I. Perrett. Time course of neural responses discriminating different views of the face and head. *Journal of Neurophysiology*, 68(1):70–84, 1992.

M. W. Oram and D. I. Perrett. Modeling visual recognition from neurobiological constraints. *Neural Networks*, 7(6/7):945–972, 1994.

M. W. Oram and D. I. Perrett. Integration of form and motion in the anterior superior polysensory area (stpa) of the macaque monkey. *Journal of Neurophysiology*, 76:109–129, 1996.

M. W. Oram, M. C. Wiener, R. Lestienne, and B. J. Richmond. Stochastic nature of precisely timed spike patterns in visual system neural responses. *Journal of Neurophysiology*, 81:3021–3033, 1999.

M. W. Oram, N. G. Hatsopoulos, B. J. Richmond, and J. P. Donoghue. Excess synchrony in motor cortical neurons provides redundant direction information with that from coars temporal measures. *Journal of Neurophysiology*, 86(4): 1700–1716, 2001.

M. W. Oram, D. Xiao, B. Dritschel, and K. R. Payne. The temporal resolution of neural codes: does response latency have a unique role? *Philosophical Transactions of the Royal Society of London*, 357:987–1001, 2002.

S. E. Palmer. Hierarchical structure in perceptual representation. *Cognitive Psychology*, 9:441–474, 1977.

S. E. Palmer. *Vision Science: Photons to Phenomenology.* MIT Press, 1999.

S. Panzeri, S. R. Schultz, A. Treves, and E. T. Rolls. Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society of London Series B: Biological Sciences*, 266:1001–1012, 1999.

A. J. Parker and W. T. Newsome. Sense and the single neuron: Probing the physiology of perception. *Annual Reviews of Neuroscience*, 21:227–277, 1998.

J. J. Peissig, E. A. Wasserman, M. E. Young, and I. Biederman. Learning an object from multiple views enhances its recognition in an orthogonal rotational axis in pigeons. *Vision Research*, 42:2051–2062, 2002.

A. P. Pentland. Automatic extraction of deformable part models. *International Journal of Computer Vision*, 4:107–126, 1990.

D. Perrett, M. Oram, M. Harries, R. Bevan, J. Hietanen, P. Benson, and S. Thomas. Viewer-centred and object-centred coding of heads in macaque temporal cortex. *Experimental Brain Research*, 86:159–173, 1991.

D. I. Perrett and M. W. Oram. Neurophysiology of shape processing. *Image and Vision Computing*, 11:317–333, 1993.

D. I. Perrett, E. T. Rolls, and W. Caan. Visual neurons responsive to faces in the monkey temporal cortex. *Experimental Brain Research*, 47:329–342, 1982.

D. I. Perrett, M. W. Oram, and E. Ashbridge. Evidence accumulation in cell populations responsive to faces: an account of generalisation of recognition without mental transformations. *Cognition*, 67:111–145, 1998.

G. Peters. Theories of three-dimensional object perception: A survey. In *Recent Research Developments in Pattern Recognition*. Transworld Research Network, 2000.

W. Pitts and W. McCullough. How we know universals: The perception of auditory and visual forms. *Bulletin of Mathematical Biophysics*, 9:127–147, 1947.

T. Poggio and S. Edelman. A network that learns to recognize three-dimensional objects. *Nature*, 343:263–266, 1990.

T. Poggio, V. Torre, and C. Koch. Computational vision and regularization theory. *Nature*, 317:314–319, 1985.

E. O. Postma, , H. J. den Herik, and P. T. W. Hudson. SCAN: A scalable neural model of covert attention. *Neural Networks*, 10(6):993–1015, 1997.

M. Potter. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning an Memory*, 2:509–522, 1976.

Y. Prut, E. Vaadia, H. Bergman, I. Haalman, H. Slovin, and M. Abeles. Spatiotemporal structure of cortical activity: properties and behavioral relevance. *Journal of Neurophysiology*, 79:2857–2874, 1998.

M. Quast and J. Teichert. Linear feature extraction based on edge orientation maps. In *Proceedings of IMVIP2001*, pages 236–242, Maynoth, Ireland, 2001.

H. S. Ranganath and G. Kuntimad. Image segmentation using pulse coupled neural networks. In *Proceedings Applications ans Science of Artificial Neural Networks II*, volume 2760, pages 543–554, Orlando, FL, April 1996.

R. Rao and D. Ballard. Dynamic model of visual recognition predicts neural response properties in the visual cortex. *Neural Computation*, 9:721–763, 1997.

R. P. N. Rao and D. H. Ballard. The visual cortex as a hierarchical predictor. Technical Report 96.4, National Resource Laboratory for the Study of Brain and Behavior, Department of Computer Science, University of Rochester, Rochester, NY 14627-0226, USA, 1996.

S. Ravishankar, A. Jain, and A. Mittal. Multi-stage contour based detection of deformable objects. In *In ECCV*, pages 483–496, 2008.

A. N. Redlich. Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, 5:289–304, 1993.

M. Riesenhuber and T. Poggio. Hierarchical models of object recognition in cortex. *Nat. Neuroscience*, 2(11):1019–1025, 1999a.

M. Riesenhuber and T. Poggio. Are cortical models really bound by the "binding problem"? *Neuron*, 24:87–93, 1999b.

I. Rock. *Perception*. Scientific American Library, 1985.

I. Rock and J. DiVita. A case of viewer-centered object perception. *Cognitive Psychology*, 19(2):280–293, 1987.

M. A. Rodrigues, editor. *Invariants for Pattern Recognition ans Classification*. World Scientific, 2000.

E. T. Rolls and S. M. Stringer. Invariant visual object recognition: A model, with lighting invariance. *Journal of Physiology - Paris*, 100:43–62, 2006.

W. D. Ross and E. Mingolla. Recent progress in modeling neural mechanisms of form and color vision. *Image and Vision Computing*, pages 447–472, 1998.

W. D. Ross, S. Grossberg, and E. Mingolla. Visual cortical mechanisms of perceptual grouping: interacting layers, network, columns, and maps. *Neural Networks*, 13:571–588, 2000.

D. L. Ruderman. The statistics of natural images. *Network: Computation in Neural Systems*, 5:517–548, 1994.

E. Salinas and L. F. Abbot. Invariant visual responses from attentional gain fields. *Journal of Neurophysiology*, 77:3267–3272, 1997.

S. Sarkar and K. L. Boyer. Perceptual organization in computer vision: a review and a proposal for a classificatory structure. *IEEE Transactions on Systems, Man an Cybernetics*, 23:382–399, 1993.

S. Satoh, J. Kuroiwa, H. Aso, and S. Miyake. Recognition of rotated patterns using Neocognitron. In *Proceedings of the International Conference on Neural Information Processing*, pages 1:112–116, 1997.

S. Satoh, J. Kuroiwa, H. Aso, and S. Miyake. Pattern recognition system with top-down process of mental rotation. In *Proceedings of the International Workconference on Artificial and Natural Neural Networks (IWANN)*, pages 816–825, 1999.

B. C. Schäfer. Evaluating the applicability of image structure detectors for recognition of political gestures. Diploma Thesis, University of Bremen, 2009.

T. B. Schillen and P. König. Stimulus-dependent assembly formation of oscillatory responses: II. desynchronization. *Neural Computation*, 3:167–178, 1991.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press, 2001.

C. Schmid. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *In CVPR*, pages 2169–2178, 2006.

A. Selinger and R. C. Nelson. A perceptual grouping hierarchy for appearance-based 3D object recognition. *Computer Vision and Image Understanding*, 76 (1):83–93, 1999.

T. Serre, M. Riesenhuber, J. Louie, and T. Poggio. On the role of object-specific features for real world object recognition in biological vision. In H. e. a. Bülthoff, editor, *Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pages 387–397, Berlin Heidelberg, 2002. BMCV 2002, LNCS 2525, Springer-Verlag.

T. Serre, L. Wolf, and T. Poggio. Object recognition with features inspired by visual cortex. In *In CVPR*, pages 994–1000, 2005.

D. Shen and H. S. l. Horace. Generalized affine invariant image normalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5):431–440, May 1997.

R. N. Shepard. Cognitive psychology: A review of the book by U. Neisser. *American Journal of Psychology*, 81:285–289, 1968.

R. N. Shepard and S. Chipman. Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, 1:1–17, 1970.

R. N. Shepard and J. Metzler. Mental rotation of three-dimensional objects. *Science*, 171:701–703, 1971.

W. Singer and C. M. Gray. Visual feature integration and the temporal correlation hypothesis. *Annual Reviews of Neuroscience*, 18(555-586), 1995.

M. Sonka and V. Hlavac. *Image Processing, Analysis, and Machine Vision*. Cengage Learning, 2007.

Stemmer. *The Imaging and Vision Handbook*. STEMMER IMAGING GmbH, 2011.

J. V. Stone. Learning perceptually salient visual parameters using spatiotemporal smoothness constraints. *Neural Computation*, 8(7):1463–1492, 1996.

E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Learning hierarchical models of scenes, objects, and parts. In *In IEEE Intl. Conf. on Computer Vision*, pages 1331–1338, 2005.

E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky. Describing visual scenes using transformed objects and parts. *International Journal of Computer Vision*, 77(1-3):291–330, 2008.

P. Suppes, M. Pavel, and J. Falmagne. Representations and models in psychology. *Annual Review of Psychology*, 45:517–544, 1994.

R. Szeliski. *Computer Vision: Algorithms and Applications*. Springer, Berlin, 2010.

K. Tanaka. Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, 19:109–139, 1996.

K. Tanaka. Mechanisms of visual object recognition: monkey and human studies. *Current Opinion in Neurobiology*, 7:523–529, 1997.

K. Tanaka, H. Saito, Y. Fukada, and M. Moriya. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, 66:170–189, 1991.

M. Tarr. Rotating objects to recognize them: A case study on the role of viewpoint dependency in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, 2:55–82, 1995.

M. Tarr, P. Williams, W. G. Hayward, and I. Gauthier. Three-dimensional object recognition is viewpoint dependent. *Nature Neuroscience*, 1:275–277, 1998.

J. G. Taylor. Neural networks for consciousness. *Neural Networks*, 10:1207–1226, 1997.

J. Teichert and R. Malaka. A learning algorithm for improved pattern synchronization in networks with biologically motivated neurons. In *Proceedings of the International Joint Conference on Neural Networks IJCNN*, pages 3:273–278, 2000.

J. Teichert and R. Malaka. A component association architecture for image understanding. In *Proceedings of the ICANN 2002*, Madrid, 2002.

J. Teichert and R. Malaka. An association architecture for the detection of objects with changing topologies. In *Proceedings of the International Joint Conference on Neural Networks, IJCNN 2003*, pages 125–130, Portland, USA, 2003.

J. Teichert and R. Malaka. Iterative context compilation for visual object recognition. *Proceedings of the European Symposium on Neural Networks, ESANN*, 2006.

S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381(520-522), 1996.

F. Tong. Primary visual cortex and visual awareness. *Nature*, 4:219–229, 2003.

K. Tsunoda, Y. Yamane, M. Nishizaki, and M. Tanifuji. Complex objects are represented in macaque inferotemporal cortex by the combination of feature columns. *Nature Neuroscience*, 4(8):832–838, 2001.

S. Ullman. Aligning pictoral descriptions: an approach to object recognition. *Cognition*, 32(3):193–254, 1989.

S. Ullman. *High-level vision.* MIT Press, 1995.

S. Ullman. Computation of pattern invariance in brain-like structures. *Neural Networks*, 12:1021–1036, 1999.

L. G. Ungerleider and M. Mishkin. Two cortical visual systems. In D. J. Ingle, M. A. Goodale, and R. J. W. Mansfield, editors, *Analysis of Visual Behaviour*, chapter 18, pages 549–586. MIT Press, Cambridge, MA, 1982.

E. Vaadia, I. Haalman, M. Abeles, H. Bergman, Y. Prut, H. Slovin, and A. Aertsen. Dynamics of neuronal intersections in the monkey cortex in relation to behavioral events. *Nature*, 373(9):515–518, 1995.

D. C. Van Essen and H. A. Drury. Structural and functional analysis of human cerebral cortex using a surface-based atlas. *Journal of Neuroscience*, 17(18): 7079–7102, 9 1997.

D. C. Van Essen and J. H. R. Maunsell. Hierarchical organization and functional streams in the visual cortex. *Trends in neurosciences*, 6(9):370–375, 1983.

G. J. van Tonder and Ejima. The 'patchwork engine': Image segmantation from shape symmetries. *Neural Networks*, 13(3):291–303, 2000.

V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer, 1999.

P. A. Viola and M. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

R. Vogels and I. Biederman. Effects of illumination intensity and direction on object coding in macaque inferior temporal cortex. *Cerebral Cortex*, 12:1047–3211, 2002.

R. Vogels, I. Biederman, M. Bar, and A. Lorincz. Inferior temporal neurons show greater sensitivity to nonaccidental than to metric shape differences. *Journal of Cognitive Neuroscience*, 13(4):444–453, 2001.

von der Malsburg. Network and self-organization. In S. Zornetzer, J. Davis, and C. Lau, editors, *An Introduction to Neural and Electronic Networks*, pages 421–432, San Diego, CA, 1990. Academic Press.

C. von der Malsburg. The correlation theory of brain function (reprint from 1981). In E. Domany, J. v. Hemmen, and K. Schulten, editors, *Models of neural networks II*, pages 95–119. Springer, Berlin, 1994.

C. von der Malsburg. Binding in models of perception and brain function. *Current Opinion in Neurobiology*, 5:520–526, 1995.

C. von der Malsburg and J. Buhman. Sensory segmentation with coupled neural oscillators. *Biological Cybernetics*, 67:233–242, 1992.

C. von der Malsburg and W. Schneider. A neural cocktail-party processor. *Biological Cybernetics*, 54:29–40, 1986.

J. C. Vorbrüggen and C. v.d. Malsburg. Data-driven segmentation of grey-level images with coupled nonlinear oscillators. *Proceedings ICANN'95*, 2:297–302, 10 1995.

G. Wallis and E. T. Rolls. Invariant face and object recognition in the visual system. *Progress in Neurobiology*, pages 167–194, 1997.

D. Wang, J. Buhmann, and C. von der Malsburg. Pattern segmentation in associative memory. *Neural Computation*, 2:94–106, 1991.

G. Wang, M. Tanifuji, and K. Tanaka. Functional architecture in monkey inferotemporal cortex revealed by in vivo optical imaging. *Neuroscience Research*, 32:33–46, 1998.

E. K. Warrington and A. M. Taylor. Contribution of the right parietal lobe to object recognition. *Cortex*, 9:152–164, 1973.

M. Wehr and G. Laurent. Odour encoding by temporal sequences of firing in oscillating neural assemblies. *Nature*, 384:162–166, 1996.

T. Wennekers and N. Ay. Spatial and temporal stochastic interaction in neural assemblies. *Theory in Biosciences*, 122:5–18, 2003.

W. A. Wickelgren. Context sensitive coding, associative memory and serial order in (speech) behavior. *Psychological Review*, 76:1–15, 1969.

L. Wiskott. Slow feature analysis: A theoretical analysis of optimal free responses. *Neural Computation*, 15(9):2147–2177, 2003.

L. Wiskott. How does our visual system achieve shift and size invariance. In J. L. van Hemmen and T. J. Sejnowski, editors, *Problems in Systems Neuroscience*. Oxford University Press, 2004.

L. Wiskott and T. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. *Neural Computation*, 14(4):715–770, 2002.

L. Wiskott, J. M. Fellous, N. Krüger, and C. von der Malsburg. Face recognition by elastic bunch graph matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):775–779, 1997.

R. P. Würtz. *Multilayer dynamic link networks for establishing image point correspondences and visual object recognition.* PhD thesis, Ruhr-Universität Bochum, Fakultät für Physik und Astronomie, Bochum, 12 1994.

R. P. Würtz. *Multilayer Dynamic Link Networks for establishing Image Point Correspondences and Visual Object Recognition.* Verlag Harri Deutsch, Thun, Fankfurt am Main, volume 41 of reihe physik edition, 1995.

R. P. Würtz. Neural networks as a model for visual perception: What is lacking? *Cognitive Systems*, 5(2):103–112, 1999.

R. S. Zemel and G. E. Hinton. Developing population codes by minimizing description length. *Neural Computation*, 7:549–564, 1995.

J. Zhang, M. Marszaöek, S. Lazebnik, and C. Schmid. Local features and kernals for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

Index